WHAT DOES DISTILLATION TRANSFER? CAUSAL TESTS WITH GROUND-TRUTH TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distilled students can reproduce a teacher's answers without using the same computation. We test what transfers through logits, hidden states, and state traces using modular-addition and automaton teachers with measurable internal signatures. In modular addition, a model's wrong-answer probabilities carry the Fourier frequencies it computes with, and changing only those probabilities selects which frequencies a freshly initialized student uses. Exchanging these probabilities between teachers transfers the source teacher's causal frequency signature. Permuting them while preserving entropy and the correct-answer probability selects a different signature, predicted before training from the edited target's frequency content. Removing the predicted frequencies sharply reduces accuracy; retaining them preserves accuracy in the trained network. A student that already computes through one frequency set can later match a new target and still depend on the old set. Raising the loss weight on those same wrong-answer probabilities restores transfer from near-deterministic teachers. Program-generated automaton traces teach state tracking but contain no teacher identity at training lengths. Using 1.5B teachers and 0.5B students, the probabilities on tokens other than the reference token again determine which teacher's output signature the student acquires. Swapping the output projection moves most of the measured gap. Distillation can therefore select a causal internal signature, but matching the target does not guarantee that the student computes through it.

1 INTRODUCTION

The problem. A distilled student that reproduces its teacher's outputs has learned one of three things: the teacher's mechanism, a different mechanism computing the same function, or a shortcut that works only within the training distribution. Standard benchmarks cannot distinguish these cases. Yet the difference determines whether verified properties of the teacher carry over to the student, and whether the student will generalize the same way. Frontier reasoning models are routinely distilled into smaller models by fine-tuning on their traces (Guo et al., 2025; Yue et al., 2025; Kim et al., 2025). Prior work tests whether teacher and student depend on the same input features under counterfactual changes (Wu et al., 2024) or compares circuits across a real distilled pair (Haskins & Adams, 2026). We test internal implementation transfer where teacher mechanisms are known and ordinary task behavior cannot distinguish teacher and student.

What we did. We construct teachers whose mechanisms can be verified directly. Modular-addition teachers use identifiable Fourier frequency sets (Nanda et al., 2023; Zhong et al., 2023), and independently trained teachers land on nearly disjoint sets, so each teacher carries a fingerprint. Automaton teachers carry state representations whose causal role we test by activation patching. For each pair of teachers with distinct mechanisms, we train identically initialized students using logit matching (varying temperature and KL direction), hidden-state matching, and state traces, and we vary student capacity. Each *channel* is defined by what the student is trained to match. For the logit channel, the *tail* is the distribution over wrong answers. We then compare each student with its own teacher and with independent teachers, using probes, causal interventions, and structural fingerprints. All thresholds and null-construction rules were fixed before the corresponding student results were evaluated (one preregistered expansion, computed afterward, is in Appendix N). A cell

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is one configuration (task, channel, loss settings, teacher); the initial experiments use five student seeds per cell, and later experiments state their seed counts.

What we found. 1. Output targets select causal signatures without guaranteeing computational identity. Softening near-deterministic targets transfers causally used modular-addition frequencies; reverse KL gives a partial copy and hidden-state matching the strongest. An entropy-preserving tail permutation selects a predicted new frequency set, but established students can fit a donor's tail while retaining their original frequency dependence (Sections 4 and 5). Fresh students can also match an attention-dependent teacher's tail without inheriting its accuracy dependence on its learned attention pattern (Appendix R).

2. The wrong-answer distribution and its loss weight both matter. Standard forward KL at $T{=}1$ nearly ignores those probabilities; raising their weight restores the modular-addition signature, while a uniform distribution at the same weight transfers none of it. Automaton students under the same re-weighting copy the teacher's wrong-answer distribution but do not match its hidden-state geometry. A large between-teacher KL is a useful filter, not a guarantee of an internal match (Section 6; Appendices N and P; Appendix J summarizes the channels).

3. Automaton state traces transfer procedure, not teacher identity. Both teach the same causally verified state tracker. At training lengths the traces are identical, so there is no teacher identity to inherit. Corruption reduces length generalization; internal measurements track the failure only at the longer lengths (Sections 7 and 8).

4. Capability and decodability need separate tests. Procedure-only traces raise pass@k by narrowing the answer support, not by the demonstrated composition (Appendix D), and direct-mode automaton teachers have decodable states that fail causal interchange (Section 9).

5. The same tail contrast appears at pretrained scale. A divergent 1.5B teacher pair transfers its tail signature to 0.5B students; the reference-token probability sets the shift's size, not its direction, and most of the shift survives a partial same-reference-token shuffle (Section 10).

2 Related work

Distillation and the capability boundary. Knowledge distillation (Hinton et al., 2015) originally targeted compression; its modern use transfers reasoning ability by fine-tuning on teacher traces (Guo et al., 2025). Yue et al. (2025) motivate this line of work: reinforcement learning with verifiable rewards sharpens what a base model can already sample, while distillation expands the pass@k boundary. Kim et al. (2025) split traces behaviorally into reasoning patterns and domain knowledge, and conjectured that only new knowledge expands capability; Appendix D tests that conjecture where both are constructed rather than inferred. On-policy variants (Agarwal et al., 2024) and trace curation (Griot et al., 2026) modify the trace channel we study.

Soft targets and what they buy. The classical account credits distillation's gains to dark knowledge (Hinton et al., 2015); later work complicated it (Furlanello et al., 2018; Phuong & Lampert, 2019; Stanton et al., 2021), but always with agreement or accuracy as the dependent variable. Zhao et al. (2022) decompose the KD loss into target-class and non-target-class terms and credit the gains to the latter, again measured as accuracy. When we saturate behavior in every cell, the soft distribution's effect shows up as the mechanism itself: the tail-swap experiment makes the non-target term the carrier of a teacher-specific causal signature (Appendix H). We also examine two established methods adopted for output quality, reverse KL in language-model distillation (Gu et al., 2024) and hidden-state hints (Romero et al., 2015); we report what each transfers internally. Chain-of-thought faithfulness work (Turpin et al., 2023; Lanham et al., 2023) shows stated reasoning can diverge from the computation behind it. Our dual-mode teachers are the training-side analog. One set of weights produces traced and direct computations that differ under every causal criterion we apply.

What distillation does inside the network. Wu et al. (2024) test shared dependence on input features under counterfactual changes, deliberately abstracting away internal implementation. We test teacher-specific internal signatures when ordinary task behavior does not distinguish the models, and manipulate the target component that carries those signatures. Haskins & Adams (2026) study circuit reorganization in a real teacher–student pair. Wadhwa et al. (2025) explicitly impose circuit

alignment; we test standard losses without that constraint. Feature comparisons (Shi et al., 2026) and provenance tracing (Liu et al., 2026) provide correlational evidence. Our constructed teachers permit causal tests.

Trait transmission through generated data. Subliminal learning shows that students acquire a teacher's behavioral traits from semantically unrelated generated data (Cloud et al., 2026). Schrodi et al. (2026) trace the effect to the rare tokens where differently biased teachers disagree. This parallels our finding that between-teacher divergence carries the signal. The inherited object is a verified implementation manipulated directly.

Mechanism identification in small models. Teachers with readable mechanisms come from the grokking literature: modular-addition networks implement Fourier multiplication (Power et al., 2022; Nanda et al., 2023), distinct runs land on distinct solutions (Zhong et al., 2023), and the frequency set is a per-model fingerprint. Linear probes (Alain & Bengio, 2016) and activation patching (Meng et al., 2022; Geiger et al., 2021) supply the identity criteria. Our dissociation reinforces the case that probes alone overclaim (Belinkov, 2022).

3 SETTING: TEACHERS WITH VERIFIED MECHANISMS

Tasks. We study two tasks, each with a different way to measure mechanism identity. In modular addition, the model reads $a + b =$ and predicts $a + b \bmod 97$ at the final position. All 97^2 pairs are split $50/50$ into train and held-out sets. Generalization is composition of unseen operand pairs. A trained network solves this task by rotation: each residue is embedded as a point on a circle, adding operands becomes adding angles, and the answer is read off by matching the summed angle against the 97 residues (Nanda et al., 2023). A *frequency* f is a circle completing f laps as the residue runs from 0 to 96; any frequency works, a model uses several, and which set it settles on is an accident of initialization, so the set is a per-model fingerprint. Because the answer is a rotation, a model's near-misses fall at offsets from the correct answer that repeat every $97/f$ positions: its wrong-answer probabilities carry its frequencies. In DFA state tracking, models read a symbol string and report the final state, trained at lengths 4–16 and evaluated to length 48, so generalization means length extrapolation. The automata have four states over a three-symbol alphabet, and are either an absorbing-sink machine whose answer is often decided early or a parity-pair machine where every symbol matters. DFA teachers are dual-mode, trained both to answer directly and to emit the state sequence, so trace and logit channels draw on the same frozen teacher.

Identity criteria. We say a student and teacher share a mechanism when they compute the task the same way. No single measurement establishes this, so we report every criterion separately, alongside a preregistered summary classification. A1, probe correspondence: held-out linear probes recover the teacher's own intermediate variables from the student (its key Fourier frequencies of $a+b$; automaton state), at accuracies comparable to the teacher's. A2, causal interchange: patching the mechanism-carrying representation within a model produces the counterfactual output the mechanism predicts, while a mismatched-location patch does not; A2 certifies a load-bearing representation, not whose. A3, structural signature: for modular arithmetic, the *key-frequency set*, the frequencies holding at least 5% of residue-embedding Fourier power; for automata, transition-matrix recovery from probed states plus a linear centered kernel alignment (CKA) fingerprint. A4, failure correspondence, compares error patterns and applies only to models that make enough errors to score; the verdicts below use A1–A3. These criteria define mechanism identity only for the constructed tasks studied here. We make no claim to a universal definition of algorithmic equivalence, and reserve *algorithm* for the Fourier constructions the grokking literature names that way. We report five-seed ranges rather than parametric intervals. Inference rests on the preregistered criterion of exceeding the 95th percentile of the independent-teacher null fixed before any student verdict. Except where noted, the main results exceed that threshold in every seed.

Teachers and null distributions. Twenty modular-addition teachers (one per seed) all reach held-out accuracy ≥ 0.999, pass A1–A3 (A4 is unscored at ceiling), and land on distinct key-frequency sets. Pairwise fingerprint overlap between independent teachers has mean 0.099, 95th percentile 0.25, maximum 0.36: the null any inheritance claim must clear. We select seeds 4 and 8 as the distinct-mechanism pair *because* their frequency sets are disjoint, which makes the logit-channel

result a controlled two-teacher case study rather than a sample over teachers. Their correct-answer logit, as a function of the operand pair, is almost entirely explained by those frequencies. Eight DFA teachers per automaton pass a per-length stability screen. For the parity machine the verified regime is traced operation (length-48 accuracy 0.71–0.99 across the eight teachers, mean 0.92), where the CKA null across independent teachers has mean 0.72 and 95th percentile 0.86.

4 THE LOGIT CHANNEL: THE TAIL OF THE TARGET SELECTS THE SIGNATURE

We vary temperature and KL direction, then edit the tail of the target. The edited tail selects a predicted frequency set in a fresh student (an already-trained student can match the new tail and still use its original set). Students share the teacher's architecture and train from fresh initializations on the teacher's train split. The objective is the standard tempered one (Hinton et al., 2015): with teacher and student answer-position logits z_t, z_s and softmax σ,

$$\mathcal{L}_{\mathrm{KD}} = T^2 \,\mathrm{KL}(\sigma(z_t/T) \,\|\, \sigma(z_s/T)) \,. \tag{1}$$

Equation 1 is the forward direction, weighting the loss by the teacher's probabilities; the reverse variant exchanges the arguments and weights by the student's. In the initial logit-KD experiments (Table 1), every student reaches held-out accuracy at least 0.999 and agrees with its teacher's argmax on at least 99.9% of held-out pairs. Behaviorally they are indistinguishable. Internally they span the full range of outcomes. Whether the soft distribution matters for student accuracy is well studied; we ask which mechanism the student acquires when accuracy is saturated either way.

task	channel	configuration	held-out acc.	overlap own	overlap other
mod add	logit, fwd KL	$T{=}1$	0.999–1.000	0.00–0.33	0.00–0.23
mod add	logit, rev KL	$T{=}1$	1.000	0.33–0.70	0.00–0.07
mod add	logit, fwd KL	$T{=}2, 4, 8$	1.000	0.75–1.00	0.00
mod add	hidden-state	relative MSE + labels	1.000	0.88–1.00	0.00
mod add	logit, fwd KL	$T{=}4$, width 256	1.000	0.86–1.00	0.00
mod add	logit, fwd KL	$T{=}4$, width 64	1.000	0.75–0.86	0.00
mod add	logit, fwd KL	$T{=}4$, width 32	1.000	0.42–0.67	0.00
independent-teacher null (p95)				0.25	

Table 1: Modular-addition distillation. Each row is the min–max range across students. Overlap is the fraction of key frequencies in the union of the student and teacher sets that the two share.

At $T{=}1$ the forward-KL target is nearly one-hot. The verified teachers are nearly deterministic, or *sharp* (mean output entropy 0.002 and 0.009 nats), and their argmax equals the true label, so at temperature 1 the forward-KL target is within rounding of one-hot. The ten students trained this way look like independently trained models: own-teacher fingerprint overlaps of 0.00–0.33, within the null's range. Seven of the ten earn the *non-teacher implementation* verdict and three are scored mixed, each on a different single criterion (Table 5). A student can agree with its teacher on every answer while sharing none of its measured implementation signature.

Softening the distribution switches inheritance on. Students of the same teachers distilled at temperature 4 inherit their teacher's implementation. Fingerprint overlap is 0.78–1.00, far above the null maximum. Causal interchange is 1.00. Nine of the ten students pass every criterion; the tenth clears interchange and signature but falls below the probe threshold (A1 value 0.73, against 0.97–0.99 for the rest) and is scored mixed. The copy is seed-invariant (all five seeds of the seed-8 cell converge to the same six of the teacher's seven frequencies) and holds at $T{=}2$, 4 and 8. Section 6 locates the onset with finer steps. Reverse KL is a second route to the tail: at $T{=}1$ it transfers the fingerprint (overlap 0.33–0.70, above the null 95th percentile, other-teacher overlap near zero) but leaves probe accuracies below teacher level, a partial transfer.

The inherited set is causally load-bearing and teacher-specific. Projecting an entire key-frequency set out of the residue embeddings costs a teacher 0.92–0.93 held-out accuracy for its own set and nothing for the other teacher's or size-matched random sets. Every $T{=}4$ student shows

the same dependence as its teacher, losing 0.93–0.98 when its own teacher's set is removed and zero when the others are. $T{=}1$ students show only patchy, seed-idiosyncratic sensitivity: own-set, other-set, and random-set damage sit in overlapping ranges (0.00–0.33, 0.00–0.21, 0.00–0.23), with nothing donor-specific (Table 17; Appendix L). No single frequency carries the effect (worst single-frequency ablation 0.27 against 0.93 for that teacher's full set), so the set is the causal unit. Softened students compute through their teacher's spectrum.

The tail selects the signature. Holding the correct-answer probability at one teacher's value and replacing only the tail, a student given teacher A's answer mass and teacher B's tail inherits *B*: tail-donor overlap 0.86–1.00 and mass-donor overlap 0.00 in all ten students across both orientations, and in a seed-matched rerun with weights kept, ablating the tail donor's input-embedding frequencies costs 0.92–0.99 held-out accuracy against 0.00 for the mass donor's set (Appendix H). A uniform tail of the same mass transfers nothing to either (cell means 0.03–0.12), and accuracy stays at 1.000. The relative tails are nearly input-invariant, so shuffling them barely changes the target and does not test dependence on input correspondence.

Predicting a new frequency set. As a function of the offset from the correct answer, the tail has a Fourier spectrum that peaks at the teacher's frequencies. The student should inherit those frequencies, so moving the peaks should move the student's set. The remapped target puts, at offset o, the source tail's probability at offset $ro \bmod p$. This leaves the correct-answer probability, the entropy, and the probability values unchanged, and each frequency f moves to rf. Students acquire the predicted set $\{rf\}$, not the source teacher's. Overlap with $\{rf\}$ is 0.86–1.00. Removing it costs 0.94–0.99 held-out accuracy; removing the source set costs 0.00–0.01; keeping only that subspace of the input embeddings leaves accuracy at 1.000 in all twelve students. All twelve reach accuracy 1.000 (Figure 1; Appendix H).

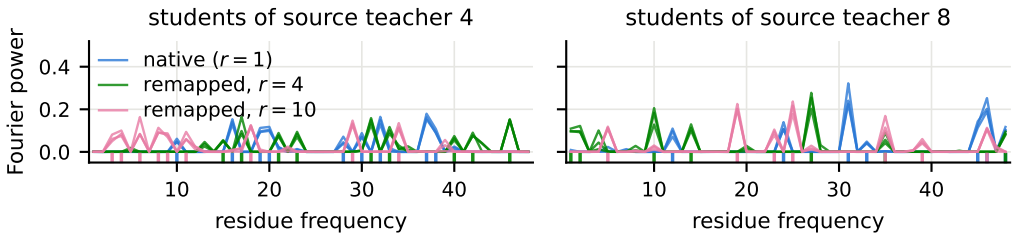


Figure 1: Embedding Fourier power of students trained on the native ($r{=}1$) and remapped ($r{=}4$, 10) tails of each source teacher, every seed drawn; peaks are the frequencies a student's embeddings use, and ticks below each axis mark the predicted set $\{rf\}$.

The student's starting point matters. An established student given another teacher's tail for 30,000 steps emits that tail (0.001–0.004 nats) yet retains causal dependence on the frequency set it already uses, and the imitation fades under gold labels. Under the remapped target we continue from two checkpoints of the same run: the earliest saved point that already depends on its current frequency set (*young*), and the 30,000-step endpoint (*mature*). After matched 30,000-step continuations, five of six young students re-select the predicted set despite that prior dependence; none of six mature students does, although five attain comparably low target KL (Appendix H).

Fresh students can also dissociate output matching from attention dependence. In an exploratory extension across two teacher pairs, all twelve students given an attention-dependent teacher's tail match it at KL 0.001–0.002 and acquire donor-associated frequency structure. Forcing attention uniform leaves their accuracy at 0.982–1.000, versus 0.379 and 0.567 for those teachers. The students' tails nevertheless shift by 0.06–0.21 nats: answer accuracy survives, but near-exact tail matching does not. This establishes differing intervention responses, not distinct algorithm identities (Appendix R).

Teacher entropy predicts $T{=}1$ transfer. Teacher 12, selected after measuring teacher entropy (0.176 nats), transfers at $T{=}1$ in all five seeds (overlap 0.78). Across ten further teachers in five prespecified pairs, all transfer at $T{=}4$ (overlap 0.33–1.00; null p95 0.25), but only four at $T{=}1$. Post hoc, entropy predicts these four (Spearman 0.879): their entropies are at least 0.0101 nats, against at most 0.0037 for the others (Appendix M). Entropy does not determine the signature: native and remapped targets select different sets at identical entropy. Appendix O reports an observational within-run checkpoint study.

The contrast repeats at 16.9M parameters. Both teachers transfer at $T{=}4$ in every seed. At $T{=}1$, the preregistered prediction holds for the higher-entropy teacher; the other transfers in two of three seeds. Size, architecture, training duration, and teacher pair change together, so this does not isolate scale (Appendix N).

Student capacity. At $T{=}4$, reducing the student's hidden dimension lowers fingerprint overlap and increases seed variation, while held-out accuracy remains 1.0 (Table 1; Figure 6, Appendix A).

5 The hidden-state channel: the strongest copies

The student matches the teacher's residual streams by mean squared error, with hard labels as the only output supervision. This is the highest-fidelity channel we test: all five of teacher 4's students land on its exact frequency set (overlap 1.00) and teacher 8's on 0.88–1.00, with zero overlap toward the other teacher. Labels alone leave students inside the independent-teacher null, so the transfer comes from the hidden term. That term must be normalized per layer by the teacher's residual power, or it trains nothing (Appendix A). The objective trains the student to match the same residual streams the identity criteria later read, so a high score is partly built in. We read this channel as an upper bound on transfer and do not rank it against the output channels.

6 Tail content and training weight jointly control inheritance

When transfer fails, does the target lack a teacher-specific signal, or does the loss give that signal too little weight? To separate channel from task, we run logit distillation on the sink-automaton teachers (verified causal state trackers in direct mode) in three cells of five students: $T{=}1$ from one teacher and $T{=}4$ from both teachers.

If temperature alone governed inheritance, $T{=}4$ should transfer a teacher-specific signature here too. It does not. All fifteen students become perfect state trackers, at ceiling accuracy through three times the training length, with interchange at 0.999; none inherits its teacher's representational geometry. CKA between a student and its own teacher spans 0.05–0.46 against an independent-teacher null with 95th percentile 0.86, and no cell meets the own-teacher match criterion.

The tasks differ in how much teacher-specific information their softened outputs carry. At $T{=}4$, the modular-addition teachers differ by a mean 0.77 nats of KL per input, the automaton teachers by only 0.0066 nats, because a four-state answer distribution at ceiling has almost no room to differ. Under these forward-KL losses, modular-addition students inherit the measured frequency signature; automaton students do not meet the CKA match criterion. Across temperatures (Table 2), modular-addition transfer begins between 0.002 and 0.009 nats and saturates by $T{=}2$, while the automaton signal never exceeds 0.007 nats and does not meet the geometry-match criterion. Temperature changes how much teacher-specific structure the output exposes. At $T{=}8$ divergence *falls* to 0.31 nats as the distributions flatten, and transfer persists anyway. Divergence is necessary here but not sufficient, as the smoothing control below shows.

The automaton null has two possible causes. Either the output target is too weak, or the only teacher-specific object is hidden-state geometry and the output carries nothing to inherit. The re-weighting experiment below separates these possibilities.

Label smoothing creates high entropy (0.79–0.80 nats against 0.025 at $T{=}1.2$) and a between-teacher divergence of 0.0099 nats, slightly above the first transferring modular-addition condition. Yet both teachers' students collapse to the same diffuse spectrum, no closer to their own teacher

task	T	divergence (nats)	identity score	null p95	transfers
mod add	1	0.0018	0.00–0.33	0.25	no
mod add	1.2	0.0093	0.50–0.86	0.25	yes
mod add	1.4	0.0302	0.56–0.86	0.25	yes
mod add	1.7	0.1027	0.75–0.86	0.25	yes
mod add	2	0.2310	0.75–0.86	0.25	yes
mod add	4	0.7689	0.78–1.00	0.25	yes
mod add	8	0.3059	0.86	0.25	yes
mod add, smoothed	1	0.0099	0.00	0.25	no
automaton	1	0.0000	0.05–0.17	0.86	no
automaton	4	0.0066	0.05–0.46	0.86	no

Table 2: Divergence vs. signature transfer, forward-KL cells. Score is frequency overlap for modular addition and CKA for the automaton. A single value means every seed agreed. The smoothed row has no key frequencies.

than to the other. The added mass is a nearly uniform tail identical for both teachers. So the residual divergence is label-confidence noise, and entropy alone carries no teacher-specific signal.

Forward KL weights the conditional tail by the teacher's wrong-answer mass, between 10^{-4} and 10^{-3} at $T{=}1$. Keeping the $T{=}1$ tail fixed and weighting its KL by $\lambda = 0.1$ gives donor key-set overlap 0.78–0.89 and own-set damage 0.47–0.92 at accuracy 1.000. Five of six uniform-tail seeds at the same weight reach accuracy 0.998–1.000 with zero donor overlap and donor-set damage of at most 0.003 (Appendix P). The sharp teacher's tail carries a signature that standard forward KL under-weights. Temperature scaling is an invertible re-encoding of a fully observed vector. It does not add information. It does change how much of that vector a given loss, precision, and budget actually use. The same re-weighting on the sink automaton splits what the temperature sweep mixed together: students match the teacher's wrong-answer distribution, and they still fail the hidden-state match. The teachers' $T{=}1$ conditional tails differ by 0.21 and 0.50 nats, more than thirty times the divergence of their softened full distributions at $T{=}4$. Re-weighted students at $\lambda \in \{0.1, 1\}$ match their own teacher's held-out tail (0.0002–0.0012 nats, versus 0.21–0.54 to the other teacher). Accuracy is 0.997–1.000 and interchange is 0.999. None meets the preregistered own-teacher CKA threshold (own 0.10–0.76, other 0.07–0.62, null 95th percentile 0.86; Appendix P).

7 The trace channel: procedure without identity

Trace students train by next-token prediction on state positions of interleaved symbol/state sequences, the small-scale analog of fine-tuning on reasoning traces. Teacher-trace cells use greedy rollouts from a frozen dual-mode teacher, errors included. The control is *gold* traces in the same format, generated by the automaton itself, correct by construction and containing no teacher. Students never see the direct question format. All measurement runs in traced mode, with its own preregistered identity criteria and nulls.

Traces teach the demonstrated procedure. All fifteen students across two teacher-trace cells and the gold cell become causally verified state trackers. Probes decode the state at symbol positions at 1.0, and decoded dynamics follow the true transition table. Patching a mid-sequence residual makes the next emitted state follow the source trace at 0.97–1.00 against a 0.0 control. These tests are not saturated: the corrupted-trace students below and the direct-mode teachers of Section 9 fail them. For the sink automaton, the learned procedure even transfers to the never-seen direct format at ≥ 0.98 accuracy for four of five seeds. The students learned a portable state machine, not a sequence habit.

No teacher identity transfers. Gold-trace students are indistinguishable from teacher-trace students under all measured criteria. Neither teacher-trace cell shows own-teacher preference: teacher 0's students have CKA 0.21–0.76 to teacher 0 and 0.21–0.82 to teacher 1, below the 0.86 null threshold. At training lengths both teachers emit the same state sequences as the gold generator, so teacher identification is at chance (0.500). Identity appears only at longer lengths, where their errors differ (Appendix K). At training lengths the traces teach the procedure and carry no teacher identity.

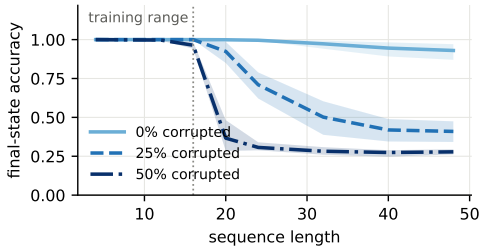


Figure 2: Final-state accuracy by length at each corruption dose; bands span five seeds.

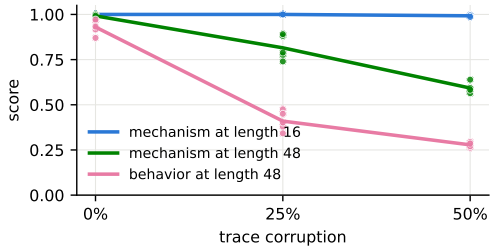


Figure 3: Same students: "mechanism" is probed-state transition consistency at lengths 16 and 48; "behavior" is final-state accuracy at 48.

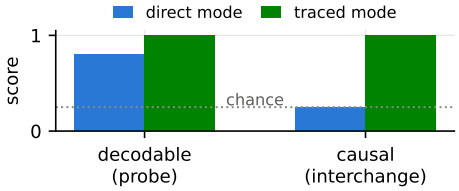


Figure 4: Eight parity-machine teachers in direct and traced modes.

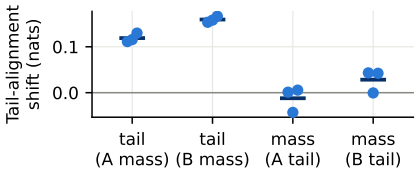


Figure 5: Single-factor donor contrasts in Δ (signed) on unscreened problems, three seeds each; bars mark means. Labels name the exchanged factor, with the fixed factor in parentheses.

8 Mechanism tracks generalization only where it is measured

Corrupting a fraction of intermediate state tokens in the gold traces (final answers kept correct) gives a controlled dose of trace corruption. Length-48 accuracy falls from 0.93 to 0.41 to 0.28 across corruption $0/25\%/50\%$, while in-range accuracy stays near ceiling (Figure 2). The internal measurements explain the collapse only when taken at the right lengths (Figure 3). At the training length, corrupted students match clean ones on every internal criterion. Probes, transition consistency, and interchange are all at ceiling even at 50% corruption, so the learned circuitry is intact where the data supported it. At length 48, probe accuracy and transition consistency both fall, tracking the behavioral collapse (Table 6, Appendix C). Internal evidence tracks capability only when measured at the distribution in question. In-range scores certify what the model *has*, not where it will keep working.

9 Decodable does not mean causally used

The parity teachers show why decodability is not enough to establish that a mechanism is used (Figure 4). When the model answers with only the final state (direct mode), probes recover the automaton state with high accuracy, and the recovered transitions match the true automaton. Every correlational criterion certifies a state-tracking mechanism. The causal test disagrees. Patching the state-carrying representation moves the output at chance, while a positive control confirms the patch itself works. The tested state representation is decodable but does not control direct-mode answers under this intervention, and the models collapse to chance immediately past the training length. When the same weights emit the state sequence (traced mode), they use the state causally, passing interchange at 0.999 and generalizing threefold. The prefix state is linearly present in direct mode because it is present in the prefix. Probes measure presence, and presence is not use.

10 PRETRAINED-SCALE EXPORT: THE TAIL SETS THE OUTPUT SIGNATURE

Do the probabilities on tokens other than the reference also determine which teacher's output signature a pretrained student acquires? Each position of a teacher-forced reference solution has two factors: the probability on that reference token and the tail. The tail is the remaining probability, renormalized over the other tokens. On GSM8K a non-reference token can begin a valid alternative solution.

The preregistered export gate requires two teachers that differ by more than retraining noise before any student trains. Short LoRA fine-tuning from a shared base fails this requirement. Further training recipes were tried with the same split, probe set, and thresholds. Students were distilled only from the one training that cleared (Table 24, Appendix Q). The tail contrast also replicates in an independently trained pair that narrowly misses the base-ratio cutoff.

We distill students from that pair: Qwen2.5-0.5B fine-tuned from its pretrained base, three seeds per target. Every identity measurement below compares 0.5B students with 1.5B teachers. Targets are built at $T{=}2$, where the gate cleared, and named by mass donor then tail donor: AA and BB are the teachers' own distributions, and AB and BA exchange the tail. Controls use uniform or same-reference-token-shuffled donor tails, and a gold-labels arm anchors the baseline.

Identity is read from Δ, the median per-token KL from the student's non-reference tail to teacher A's minus the same to teacher B's, with all models given the gold prefix on held-out problems. Student greedy problem accuracy is 0.14–0.26, a measurement separate from identity. Raw Δ carries an offset unrelated to transfer, because teacher B is about twice as high-entropy as A, so flat-tailed students read as B-like. Single-factor contrasts remove it. On unscreened problems, exchanging only the tail donor shifts Δ toward the new donor in every seed: by $+0.111$ to $+0.130$ nats with the mass at A's (AB against AA) and $+0.153$ to $+0.167$ with the mass at B's (BB against BA). Exchanging only the mass donor moves it by at most 0.043, with inconsistent sign (BA against AA, BB against AB; Figure 5). The two tail effects also differ. The exchange moves Δ further under B's mass than under A's, a paired per-problem interaction of $+0.02$ to $+0.04$ nats whose bootstrap interval excludes zero in every seed. Under plain forward KL the reference-token mass sets the tail's training weight, a coupling decoupled knowledge distillation removes (Zhao et al., 2022). This account is plausible, not demonstrated. The tail donor sets the shift's direction. The mass donor modulates its strength. A scale-free read, the cosine between mean-centered tail logits, cannot see flatness and gives the same contrasts. When donor tails are shuffled across problems with the same reference token (73% of positions), 63–75% of the shift from the gold-labels baseline remains.

Most of the inherited output signature lives in the output projection. AA and AB students share data, order, and reference-token mass donor. Copying the student's tied embedding into a separate output head leaves the logits unchanged and gives three components. Swapping the output projection alone moves 0.74–0.79 of the pure tail gap toward its donor, the decoder stack 0.18–0.20, the embedding nothing, and the two together the whole gap, at similar measured token and problem accuracy. Same-arm, cross-seed swaps move at most 0.02. The full weight difference gives either teacher's signature to a gold-labels student. Under this 600-step schedule the signature is a readout-level shift; the interventions identify no inherited computation (Appendix G reports the public-pair bridge).

11 CONCLUSION

The wrong-answer distribution can select a causal internal signature. In modular addition, swapping or remapping those probabilities chooses which frequency set a fresh student uses, and raising their loss weight restores transfer from near-deterministic teachers. Matching the target does not mean the student computes through it. An established student can match a new wrong-answer distribution and keep its old frequencies; automaton students can copy that distribution and miss the teacher's hidden-state geometry. Automaton state traces teach a procedure, not a teacher's identity, and internal scores certify the mechanism only where they are measured. At 1.5B to 0.5B the probabilities on tokens other than the reference again set the output signature, mostly in the output projection, without identifying a specific inherited computation.

Reproducibility statement

Every experiment is reproducible from the code artifact submitted as supplementary material: task generators, teacher training, the identity criteria with their unit tests, and the per-cell result JSONs behind every figure and table. Every figure and table regenerates from those JSONs alone (`./reproduce.sh figures`), and every number in the text is backed by a tracked result file.

AI use statement

An LLM-based assistant was used for portions of this work, in every case directed and reviewed by the authors. AI was used to implement methods, writing the task, evaluation, intervention, and probing code layered on the authors' transformer training library, which the authors wrote, and launching and monitoring runs; to generate the synthetic datasets, through generator programs written to the authors' specification, which emit the data in final form with no separate cleaning or reformatting step; to design and give feedback on methodology and experiments, including drafts of the preregistration documents the authors fixed before each run; to develop the conceptual framework of channels and identity criteria; to propose and refine hypotheses; to formulate the paper's two elementary mathematical claims, the standard forward-KL decomposition behind the tail re-weighting experiment, restated from prior work, and the frequency mapping of the tail permutation behind the remapping experiment, each verified by a unit test; and to interpret results. AI also drafted and revised portions of the text, edited code, produced the figures through plotting scripts, and found and formatted references. The paper contains no proofs, no translation, and no qualitative or thematic analysis, so AI was used for none of those. The authors made all final decisions on hypotheses, experimental design, preregistration, interpretation, and the paper's claims, and reviewed all AI-assisted work: code was read and exercised by tests, every reported number was checked against the result files the experiments wrote, predictions and falsifiers for the preregistered experiments were fixed in writing before the corresponding runs, and every citation was verified against its primary source. The authors take full responsibility for all content of this paper.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652:615–621, 2026.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, 2018.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Maxime Griot, Paul Steven Scotti, and Tanishq Mathew Abraham. Compress-Distill: Reasoning trace compression for efficient knowledge distillation. *arXiv preprint arXiv:2606.05988*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie

Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Reilly Haskins and Benjamin Adams. Distilled circuits: A mechanistic study of internal restructuring in knowledge distillation. *Transactions on Machine Learning Research*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith Ross. Reinforcement learning vs. distillation: Understanding accuracy and capability in LLM reasoning. *arXiv preprint arXiv:2505.14216v2*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Kaiyuan Liu, Shaotian Yan, Rui Miao, Bing Wang, Chen Shen, Jun Zhang, and Jieping Ye. Where did this sentence come from? tracing provenance in LLM reasoning distillation. In *International Conference on Learning Representations*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, 2019.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *International Conference on Learning Representations*, 2026.

Dan Shi, Zhuowen Han, Simon Ostermann, Renren Jin, Josef van Genabith, and Deyi Xiong. Why does reinforcement learning generalize? a feature-level mechanistic study of post-training in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Circuit distillation. *arXiv preprint arXiv:2509.25002*, 2025.

Cindy Wu, Ekdeep Singh Lubana, Bruno Kacper Mlodozeniec, Robert Kirk, and David Krueger. What mechanisms does knowledge distillation distill? In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, 2024.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, 2023.

A Experimental details

Tasks. The constructed-task experiments use modular addition modulo 97 and DFA state tracking. Modular-addition models train on half of the 97^2 operand pairs and are scored on the held-out half. Automaton models train at lengths 4–16 and are scored at lengths up to 48. The 16.9M-parameter pair and the 1.5B-to-0.5B export use different architectures (Appendices N and Q).

Models. Constructed-task networks are decoder-only transformers from a common implementation, with RoPE positions unless stated, RMSNorm, and a SwiGLU feedforward. The default size is two layers, hidden dimension 128, four heads, and feedforward width 512. Teachers and students in an experiment share that architecture except in the width sweep below.

Training. Table 3 lists the default configurations. Students train under the same optimizer, learning rate, weight decay, and step budget as the teachers, so a teacher–student difference is never a difference of training recipe. We did not vary weight decay or initialization scale, so we do not know how much either changes which frequency set or state representation a model settles on. Seeds control initialization and data order. Students in the initial experiments use seeds 100–104; modular-addition teachers use 0–19 and automaton teachers 0–7. Later experiments state their own seed counts.

Dual-mode automaton teachers. The two modes are separate sequence formats, with no mode token. Direct mode is the start-state token, the symbols, and an equals token, with the final state predicted at the equals position. Traced mode interleaves each symbol with the state after it, and the loss falls on every state position. A dual-mode training step sums one direct-mode batch and one traced-mode batch of 512 sequences each, so the two objectives are weighted equally at every step.

	modular addition	DFA state tracking
architecture	RoPE, RMSNorm, SwiGLU	same
layers / width / heads / FFN	2 / 128 / 4 / 512	2 / 128 / 4 / 512
optimizer	AdamW, lr 0.001, betas (0.9, 0.98)	AdamW, lr 0.001, betas (0.9, 0.98)
weight decay	1.0	0.1
steps	30,000 (full batch)	20,000 (batch 512, streamed)
warmup	10	100
train data	50% of operand pairs, $p = 97$	lengths 4–16
evaluation	held-out pairs	lengths 4, 8, 12, 16, 20, 24, 32, 40, 48

Table 3: Default constructed-task training. The DFA evaluation grid extends to three times the maximum training length.

Capacity. The width sweep is the one architectural exception. Student hidden dimension is 32, 64, or 256, and feedforward width stays four times that dimension. Figure 6 plots fingerprint overlap for those cells against the default width-128 teacher.

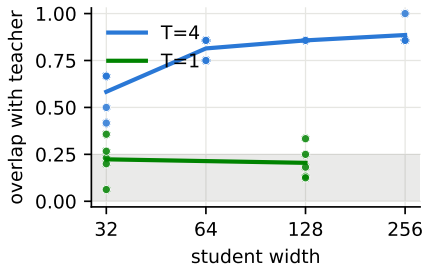


Figure 6: Fingerprint overlap against student width, teacher width 128. The same cells are the last three rows of Table 1.

Hidden-state matching. The hidden-state channel adds a per-layer residual mean squared error, divided by the teacher's residual signal power: with $h_s^{(\ell)}$ and $h_t^{(\ell)}$ the student's and teacher's residual streams after block ℓ over a training batch, the loss is the label cross-entropy plus $\frac{1}{L}\sum_\ell \mathrm{mean}\big((h_s^{(\ell)} - h_t^{(\ell)})^2\big)/\mathrm{mean}\big((h_t^{(\ell)})^2\big)$, means taken over positions and features, with unit weight on the hidden term. Modular-addition teachers train under weight decay 1.0, and their residual mean squares are 10^{-5} to 10^{-4}. Without that normalization the hidden term is negligible against the label loss and trains nothing. We confirmed the unnormalized case before switching.

Preregistration. The identity-criteria definitions, thresholds, and the independent-teacher null table were fixed before any student verdict was computed, following the project plan. Both documents ship in the code artifact's `preregistration/` folder (`battery_definitions.md` and `distillation_mechanism_transfer_plan.md`), as do the stage-level preregistrations: `stage3` (the capability factorial), `stage4` (the bridge), `stage5` (the stopped scale pilot), `stage7` (the teacher-pair replication), `stage8` (16.9M parameters), `stage9` (the checkpoint study) and `stage10` (the tail swap), each as `stageN_preregistration.md` (the bridge as `stage4_bridge_preregistration.md`), together with the stage-13 and stage-14 design records `stage13_design.md` and `stage14_design.md`. Amendments carry their own dates; none altered a threshold after the corresponding data existed, except the preregistered stage-8 null expansion (Appendix N).

B SCORING THE IDENTITY CRITERIA

A1, probe correspondence. Ridge probes ($\lambda = 10^{-2}$, bias unpenalized) are fit on a random half of the evaluation set and scored on the other half.

- Modular addition. The targets are $\cos$ and $\sin$ of each frequency of $(a+b) \bmod p$ at the answer position, scored as mean held-out R^2 per frequency. The student must reach at least 0.9 times the teacher's own value on the teacher's key set.
- Automata. The targets are one-hot states, scored as held-out accuracy per layer and position.

A2, causal interchange.

- Modular addition. The residual stream at the answer position is patched from a source example into a target example. The criterion is a counterfactual answer rate of at least 0.90 at some layer, with an operator-position control at most 0.05.
- Automata, traced mode. A mid-sequence position is patched and the next emitted state must follow the source. The patched location carries the source's state and symbol together, so the counterfactual is the source's own transition. The test shows the patched representation controls the next output, not that the model applies the transition function to a state and symbol manipulated independently.
- Automata, direct mode (Section 9). The residual stream at the best-decoding layer is patched from a source sequence into a target sequence at the position after eight of sixteen symbols. The counterfactual final state is the source's state at that point run through the target's remaining symbols by the true transition function. The control patches the start-state position, whose content is the same for every sequence. The positive control is the traced-mode test above on the same weights. Replacing the whole residual at that position leaves the final answer unchanged in 0.951–0.998 of runs across the eight teachers, so no subspace of that position, in any basis, can carry the state causally. A distributed encoding read from several positions at once remains possible and is untested. Neither automaton test carries a preregistered numerical threshold; the verdict rule applies to modular addition only, and the automaton results are reported as measured.

A3, structural signature.

- Modular addition. The signature is the key frequency set, the frequencies holding at least 5% of residue-embedding Fourier power, compared by Jaccard overlap. A match must exceed the null 95th percentile and the same student's overlap with the other teacher.
- Automata. Transition-table recovery from probed states plus a linear-CKA fingerprint at state-carrying positions of identical inputs.
- Table 4 lists the null distributions.

null distribution	pairs	mean	p95	max
Modular addition, embedding-fingerprint overlap	190	0.099	0.250	0.357
DFA decide-late, traced-mode CKA	28	0.724	0.860	0.876
DFA commit-early, direct-mode CKA	28	0.489	0.857	0.906

Table 4: Independent-teacher null distributions used by A3.

A4, failure correspondence is defined only for models with at least 50 held-out errors. It is unscored for students that receive a verdict, all of which sit at ceiling. Verdicts use the three scorable criteria.

The verdict rule. The preregistered mapping (`battery_definitions.md`, implemented in `scripts/analyze_verdicts.py`) assigns one of four labels. A student is a *teacher-signature*

match if every scorable criterion passes. It is a *non-teacher implementation* if A1 and A3 fail, A2 passes, the student decodes its own key frequencies at its causal layer with mean held-out R^2 at least 0.8, and held-out accuracy is at least 0.99. It is a *shortcut* if held-out accuracy is below 0.9 or no layer passes A2. Everything else is *partial/mixed*. Teacher 4's forward-KL student 104 shares the A1-fail, A2-pass, A3-fail pattern of students 100–103 but decodes its own key set at R^2 0.77, below the 0.8 coherence threshold, so the label is partial/mixed. Re-running the rule on the tracked criterion inputs reproduces every verdict in Table 5.

C PER-SEED RESULTS

Table 5 reports every logit-channel student's per-criterion outcome against the preregistered thresholds. Table 6 is the gold-trace corruption experiment of Section 8: accuracy, transition consistency, and probes at the training length (16) and at length 48, where accuracy collapses.

cell	student	A1	A1 val.	A2	A3	own	other	own R^2	acc.	verdict
teacher12 forward	100	✓	0.80	✓	✓	0.78	0.17	0.81	1.000	teacher-signature match
teacher12 forward	101	×	0.36	✓	✓	0.78	0.17	0.45	1.000	partial/mixed
teacher12 forward	102	×	0.55	✓	✓	0.78	0.17	0.54	1.000	partial/mixed
teacher12 forward	103	×	0.50	✓	✓	0.78	0.17	0.57	1.000	partial/mixed
teacher12 forward	104	✓	0.79	✓	✓	0.78	0.17	0.81	1.000	teacher-signature match
teacher4 forward	100	×	0.20	✓	×	0.12	0.00	0.89	1.000	non-teacher implementation
teacher4 forward	101	×	0.00	✓	×	0.00	0.23	0.97	1.000	non-teacher implementation
teacher4 forward	102	×	0.39	✓	×	0.15	0.00	0.91	1.000	non-teacher implementation
teacher4 forward	103	×	0.22	✓	×	0.12	0.07	0.98	1.000	non-teacher implementation
teacher4 forward	104	×	0.08	✓	×	0.07	0.00	0.77	0.999	partial/mixed
teacher4 forward T4	100	✓	0.97	✓	✓	0.78	0.00	0.99	1.000	teacher-signature match
teacher4 forward T4	101	✓	0.98	✓	✓	0.78	0.00	0.99	1.000	teacher-signature match
teacher4 forward T4	102	×	0.73	✓	✓	1.00	0.00	0.73	1.000	partial/mixed
teacher4 forward T4	103	✓	0.98	✓	✓	0.78	0.00	0.99	1.000	teacher-signature match
teacher4 forward T4	104	✓	0.98	✓	✓	0.89	0.00	0.99	1.000	teacher-signature match
teacher4 reverse	100	×	0.66	✓	✓	0.50	0.07	0.96	1.000	partial/mixed
teacher4 reverse	101	×	0.76	✓	✓	0.70	0.00	0.97	1.000	partial/mixed
teacher4 reverse	102	×	0.54	✓	✓	0.40	0.00	0.97	1.000	partial/mixed
teacher4 reverse	103	×	0.54	✓	✓	0.40	0.00	0.97	1.000	partial/mixed
teacher4 reverse	104	×	0.54	✓	✓	0.44	0.00	0.78	1.000	partial/mixed
teacher8 forward	100	✓	0.95	✓	×	0.18	0.07	0.96	1.000	partial/mixed
teacher8 forward	101	×	0.28	✓	×	0.13	0.19	0.88	1.000	non-teacher implementation
teacher8 forward	102	×	0.55	✓	×	0.25	0.13	0.96	1.000	non-teacher implementation
teacher8 forward	103	×	0.18	✓	×	0.12	0.18	0.80	1.000	non-teacher implementation
teacher8 forward	104	×	0.84	✓	✓	0.33	0.00	0.99	1.000	partial/mixed
teacher8 forward T4	100	✓	0.99	✓	✓	0.86	0.00	0.99	1.000	teacher-signature match
teacher8 forward T4	101	✓	0.99	✓	✓	0.86	0.00	0.99	1.000	teacher-signature match
teacher8 forward T4	102	✓	0.99	✓	✓	0.86	0.00	0.99	1.000	teacher-signature match
teacher8 forward T4	103	✓	0.99	✓	✓	0.86	0.00	0.99	1.000	teacher-signature match
teacher8 forward T4	104	✓	0.99	✓	✓	0.86	0.00	0.99	1.000	teacher-signature match
teacher8 reverse	100	×	0.56	✓	✓	0.57	0.00	0.98	1.000	partial/mixed
teacher8 reverse	101	×	0.56	✓	✓	0.50	0.00	0.97	1.000	partial/mixed
teacher8 reverse	102	×	0.42	✓	✓	0.38	0.00	0.98	1.000	partial/mixed
teacher8 reverse	103	×	0.42	✓	✓	0.33	0.00	0.91	1.000	partial/mixed
teacher8 reverse	104	×	0.69	✓	✓	0.50	0.00	0.85	1.000	partial/mixed

Table 5: Per-seed, per-criterion verdicts for the seven modular-addition logit-KD cells that carry a verdict. A1 value is mean held-out probe R^2 over the teacher's key frequencies at the student's causal layer. Own and other are fingerprint overlaps with the two teachers of the pair. Own R^2 is the student's mean held-out probe R^2 over its own key frequencies (the 0.8 coherence condition), and acc. is its held-out accuracy (the 0.99 condition). Teacher 12's other teacher is teacher 8. The displayed verdict labels rename the preregistration's terms (teacher-algorithm, different-valid-algorithm, mixed) without changing the preregistered decision rule.

corruption	acc. len 16	acc. len 48	consistency len 16	consistency len 48	probe past 32
0%	1.00	0.87–0.97	1.00	0.98–1.00	0.98–1.00
25%	1.00	0.34–0.47	1.00	0.74–0.89	0.62–0.84
50%	0.94–0.98	0.26–0.29	0.99–1.00	0.56–0.64	0.31–0.46

Table 6: Gold-trace corruption at lengths 16 and 48, five seeds per dose (ranges). Probe past 32 is held-out probe accuracy for the state after each of symbols 33–48 at the best-decoding layer of length-48 sequences; consistency is measured at the same layer.

D KNOWLEDGE AND PROCEDURE TRACES CONTROL DIFFERENT OUTCOMES

To test how the trace channel's content affects capability, we construct two synthetic knowledge graphs in which we know which facts the student lacks. The teacher knows every fact and demonstrates two-hop composition. The base student knows facts for a known 70% of the entities and, by acceptance test, recalls held-out facts at chance. We then vary two factors independently: whether the training traces contain the student's missing knowledge (K$\pm$) and whether they demonstrate the two-hop procedure (P$\pm$). A separate condition trains the same base student by group relative policy optimization (GRPO) on exact-match reward with no traces. Evaluation uses relation paths that are absent from the distillation data; GRPO trains on those paths. Each cell uses five seeds, giving ten runs across the two graphs. We score pass@k to 256. The remaining setup is in Appendix E.

Each factor controls one outcome and leaves the other alone (Figure 7). Traces carrying the missing facts teach them completely: the student recalls every held-out fact and answers single-hop questions about them at pass@256 of 1.00. It never chains two facts together to answer a two-hop question. Traces demonstrating the procedure do the reverse. They teach no new facts, yet the student answers two-hop questions along unseen relation paths, reaching 0.20–0.37 at pass@256 against 0.00 for the base student in all ten runs. Its margin over guessing an answer of the right form is widest mid-range: about five times that chance rate at k=16, four times at k=32 as the rate itself climbs.

The larger combined K+P+ dataset, trained at the same step budget, underperforms each specialist on its own job. Dilution, capacity and gradient interference are candidate causes we have not separated (Appendix E). Kim et al. (2025) conjecture that "capability improves only when new knowledge is introduced, whereas distilling reasoning patterns only improves accuracy but not capability". Facts-only traces improve held-out fact recall. Procedure-only traces introduce no new facts under the recall test, yet raise pass@k on unseen paths. That pass@k gain does not require new facts. The students also did not learn the teacher's procedure. In sampled chains that end in the correct answer, the teacher states the correct first hop every time, and the students almost never do, 0.00 in nine of ten runs and 0.24 in the tenth (Appendix E). A successful student chain typically substitutes a trained relation for the never-trained first hop, walks real facts from the wrong start, and ends on a genuine attribute read that narrows the answer support. Pass@k rises by that narrowing. GRPO from the same base gives pass@256 of 0.00–0.03 on the two-hop evaluation split (Eval-P, Appendix E) after 2,000 steps, compared with 0.00 for the base and 0.20–0.37 for procedure-trace students. Rewards were nonzero at three of 110 logged steps. The logs do not enumerate every training step. These results concern the tested sparse-reward configuration, not a general limit of reinforcement learning (Yue et al., 2025) (Appendix E).

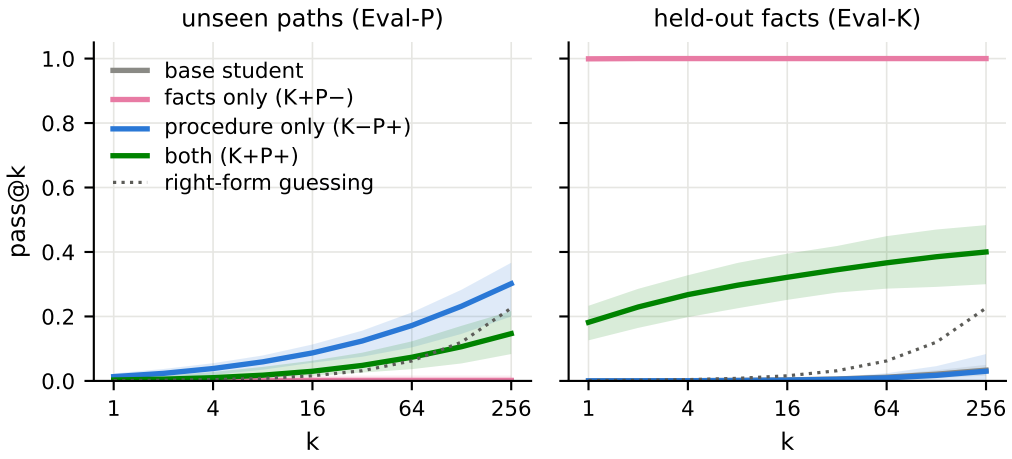


Figure 7: pass@k on two-hop questions over known facts (left) and on single-hop questions about held-out facts (right), bands over ten runs.

E KNOWLEDGE AND PROCEDURE TRACES: SETUP

The experiment uses two independently generated knowledge graphs (seeds 0 and 1). Each graph has 400 invented entities. Every entity has three attributes (city, field, year) and three relations (mentor, rival, collaborator). Attribute values are drawn from pools of 1000 per attribute, which sets the right-form guessing rate $1 - (1 - 1/1000)^k$ reported with every curve. Names and values are invented strings.

An example from the seed-0 graph, where `vexef` is in the student's known split:

Facts:

```
the city of vexef is city821.
the mentor of vexef is balolm.
```

Training question:

```
the city of the rival of the mentor of vexef?
```

Gold trace:

```
the mentor of vexef is balolm.
the rival of balolm is renent.
the city of renent is city633.  city633
```

Held-out item:

```
the city of balef?  city452
```

Training paths begin with mentor or rival. Evaluation paths begin with collaborator and never appear in training.

The teacher is trained on every fact and on worked two-hop traces. Networks have 4 layers, width 256, 4 heads, feedforward width 1024, and context 64. They train for 20,000 steps of AdamW at learning rate 3×10^{-4} over packed fact and question documents. Packing uses eight independent shuffles so no line is permanently split, and every prompt begins with the separator token. Distillation fine-tuning runs 4,000 steps at learning rate 10^{-4}. The GRPO arm runs 2,000 steps with groups of 8 over 8 questions per step, group-normalized advantages, and exact-match reward. Evaluation scores pass@k to $k = 256$ on three splits. Eval-K is single-hop questions about held-out entities. Eval-P is two-hop questions on known entities along relation paths that never appear in training. Eval-KP is two-hop questions about held-out entities on those same paths.

A teacher is accepted if single-hop recall is at least 0.99 on all entities and multi-hop accuracy is at least 0.95 on training paths. A student is accepted if recall is at least 0.99 on known entities and at most 0.15 on held-out ones. The realized students score 1.00 and 0.00.

The design and four predictions were committed before any run (`stage3_preregistration.md`).

- P1, the conjecture of Kim et al. (2025) as we read it. Pass@k on Eval-K and Eval-KP expands only in the K+ cells. Procedure-only traces raise pass@1 on Eval-P but not pass@256.
- P2, our alternative, which contradicts P1's second clause. Procedure-only traces raise Eval-P pass@256 above the base student and above traces that carry neither facts nor the procedure.
- P3. Facts-only traces raise Eval-K pass@256. They do not raise Eval-P or Eval-KP.
- P4. GRPO raises pass@1 on Eval-P. Pass@256 on every split stays within 0.05 of the base student.

Two dated amendments preceded the full run. One enlarged the value pools after a pilot hit chance on every cell. One closed a leak by which a procedure-only chain could end at a held-out entity. Table 7 reports pass@256 for every cell and split.

cell	Eval-K@256	Eval-P@256	Eval-KP@256	held-out recall
base student	0.02–0.05	0.00	0.00–0.02	0.00
K−P−	0.00–0.03	0.00–0.05	0.00–0.05	0.00
K+P−	1.00	0.00–0.02	0.00–0.03	1.00
K−P+	0.00–0.08	0.20–0.37	0.18–0.25	0.00
K+P+	0.30–0.48	0.08–0.22	0.03–0.13	0.13–0.23
RLVR (GRPO)	0.00–0.05	0.00–0.03	0.00–0.05	0.00

Table 7: Factorial pass@256 by cell and split, ranges over ten runs. Held-out recall verifies directly which cells teach the missing facts.

The GRPO arm, and what it does not show. The reward requires an exact match on the same held-out relation paths the distillation arms are evaluated on, so it fires only when the policy samples a correct multi-hop answer. Those paths are held out from the distillation arms but not from GRPO, which trains on them. A positive result under that overlap would not have been interpretable. The outcome is not positive. The training logs record the mean group reward every 200 steps. Across the ten runs these 110 logged means are zero except at three points (one run each in the seed-0 graph, student seeds 0, 2 and 4), where 1, 8 and 5 of the 64 rollouts at that step matched the answer. Groups with identical rewards contribute zero reward-based advantages, and the implementation skips the optimizer step entirely when every group's advantages vanish (the loss then has no gradient and no other term). The sparse logging does not determine the total number of training steps with nonzero advantages. After 2,000 steps the GRPO students' Eval-P pass@256 is 0.00–0.03 (at most two of sixty questions) against 0.00 for the base, and their Eval-KP and Eval-K values stay within 0.03 of the base's (Table 7). Their Eval-P pass@1 is 0.0000–0.0001 against 0.0000 for the base, so the first clause of P4 also fails. This is the near-zero-support case of the method, not a general limit on reinforcement learning. Exploration bonuses, shaped or partial-credit rewards, a curriculum over shorter paths, or a base policy with a higher initial success rate would each change the picture. We tested none of them, and the claim in Appendix D is restricted to this base, task, and GRPO configuration.

Two observations outside the predictions. Facts embedded in multi-hop traces are learned less completely than the same facts stated singly (Eval-K 0.30–0.48 versus 1.00). If the goal is to teach the facts, they belong on their own lines, not only inside traces. The combined dataset is also no better at the procedure job than the procedure-only dataset (Eval-P 0.08–0.22 versus 0.20–0.37). Those ranges are adjacent over ten runs, so we do not claim a reliable decrement on that side. Neither observation was predicted, and we have not separated the candidate causes: capacity, gradient interference between the two objectives, and the dilution just described. Procedure-only students move slightly above chance at intermediate k on questions whose facts they verifiably lack. Direct fact leakage is ruled out by held-out recall of 0.000–0.003.

The first-hop diagnostic. This check was not preregistered; it was added after the factorial's results were in (`scripts/stage3_firsthop.py`). We drew 256 fresh samples per Eval-P question at a recorded seed. For each sample we record the final answer and check whether the first hop named for the queried entity is true.

The teacher states the correct first hop in every correct-answer chain. Those chains stay true to the graph at 0.999 in both graphs. Procedure-only students state the correct first hop in 0.000 of correct-answer chains in nine of ten runs, and in 0.243 in the tenth.

Eval-P questions begin with collaborator, which never appears first in training. The students start with a trained relation instead. They then walk real facts from the wrong entity and finish by reading a real attribute of a known entity. That last read cuts down the answers that can still be right. With 256 samples, the smaller set is enough to hit the answer. Pass@k on Eval-P rises by this route. The chains do not carry out the two-hop procedure the traces showed.

F HIDDEN-STATE MATCHING ON THE AUTOMATON TASK

The main text never ran hidden-state matching on the automaton. Logit students on that task copy the tracker and miss the teacher (Section 6). This run tests whether the teacher's geometry transfers when residual streams are the training target. It uses the two teachers from the logit cells, five students per teacher.

Behavior. The students are causally verified state trackers. Interchange is 0.999 and length-48 accuracy is 1.000, the same ceiling as the logit-channel students on this task.

Geometry. Each student matches its own teacher's representational geometry. CKA is 0.994–0.999 toward its own teacher against 0.332–0.362 toward the other. The split remains at length 48, which the objective never trained on: 0.894–0.998 own against 0.194–0.272 other. An independent-teacher null at that length has mean 0.413 and maximum 0.877.

A high CKA is partly built in. The loss trains the student to match the teacher's residual streams. Teacher 0's ordinary forward-KL students at $T{=}4$ sit closer to teacher 1 in every seed (CKA 0.21–0.86 toward it against 0.05–0.43 toward their own teacher); teacher 1's sit closer to their own teacher in four of five seeds, still below the 0.86 threshold. Re-weighted logit students reproduce their own teacher's conditional tail and still miss the geometry-match criterion (Appendix P). Matching the outputs and matching the hidden-state geometry are different tests on this task.

G BRIDGE TO A REAL DISTILLED PAIR

This appendix asks whether the paper's quantities behave on a public trace-distilled model. We compare DeepSeek-R1-Distill-Qwen-1.5B against its base, Qwen2.5-Math-1.5B, on 400 GSM8K test problems, 32 samples each at temperature 0.6. No ground-truth algorithm exists at this scale, so nothing here adjudicates the mechanism used. The protocol and predictions were committed before any generation (`stage4_bridge_preregistration.md`). The predictions:

- P1 (primary). An internal readout of whether a trace ends correct should correlate with the per-problem pass@16 gain (distilled minus base) more strongly than the trace's likelihood does. The readout is a logistic probe from the mid-layer hidden state (layer 14 of 28), the coherence probe, trained on 200 problems and scored on a held-out 200. The probe reads the greedy trace, one per problem, and predicts whether that trace ends correct; the likelihood baseline is the greedy trace's mean token log-probability, so neither involves the 32 sampled solutions that define pass@16. The halfway token is the primary read; the quarter and three-quarter tokens are secondary.
- P2. The distilled model's output entropy on reasoning tokens exceeds the base model's.

The distilled model expands capability (pass@1 0.40 vs 0.27; pass@16 0.92 vs 0.89).

On P1, trace likelihood is uncorrelated with the per-problem pass@16 gain (Spearman -0.02). The probe's correlation is 0.016, 0.076, and 0.184 at the quarter, half, and three-quarter positions (Figure 8).

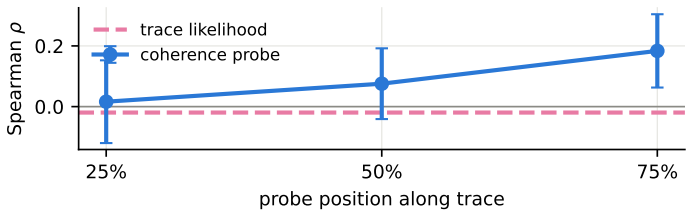


Figure 8: Coherence-probe correlation with per-problem pass@16 gain along the trace; dashed is the trace-likelihood baseline, bars are 90% bootstrap CIs on the difference.

At the primary halfway position the probe's advantage over likelihood is positive but not significant (90% bootstrap CI $[-0.02, 0.21]$). At the secondary three-quarter position it is significant ($[0.08, 0.32]$). P1 is not established at the primary position. The three-quarter read is the secondary endpoint it was registered as. Section 8 predicts this rise: internal evidence tracks capability best when read near where the computation finishes.

P2 holds: reasoning-token entropy is 0.43 nats against 0.32 for the base.

We report this as a single diagnostic on one pair and one benchmark. It checks the paper's indicators. It does not extend the mechanism claims beyond the constructed setting.

H TARGETS THAT VARY ONLY THE TAIL

Softened output targets can transfer a teacher's frequency fingerprint (Section 4). However, raising entropy is not enough. A uniform tail of the same softness transfers nothing (Section 6). Neither experiment isolates whether the identity sits in the correct-answer probability or in the distribution over wrong answers. This appendix constructs targets that hold the correct-answer probability at one teacher's value and vary only the distribution over the incorrect classes. It uses teachers 4 and 8, whose key-frequency sets are disjoint. Design and predictions were committed before any target was built (`stage10_preregistration.md`).

Tails are handled in coordinates relative to the correct answer. Entry k is the probability of $(c + k) \bmod p$, renormalized over the $p - 1$ incorrect residues. The two non-residue tokens carry about 0.9% of the mass at $T{=}4$ and pass through from the mass donor unchanged, so the intervention is confined to the residue tail. Each cell uses five student seeds and both teacher orientations, forty students in total.

tail	mass donor	overlap with mass donor	with the other teacher	accuracy
own	4	0.844	0.000	1.000
teacher 8's	4	0.000	**0.857**	1.000
own, shuffled	4	0.956	0.000	1.000
uniform	4	0.045	0.033	1.000
own	8	0.857	0.000	1.000
teacher 4's	8	0.000	**0.911**	1.000
own, shuffled	8	0.857	0.000	1.000
uniform	8	0.118	0.107	1.000

Table 8: Mean key-frequency overlap over five students per cell.

When the correct-answer probabilities come from one teacher and the tail from another, the student carries the tail donor's frequency set and none of the answer donor's. A uniform tail, with the

correct-answer mass held at the teacher's own value rather than raised, transfers nothing, matching the label-smoothing result of Section 6.

The shuffled cells were intended as a third control, permuting each teacher's own tails across inputs to destroy the input-to-tail correspondence while holding the tail entropy distribution fixed. They transfer as strongly as the unshuffled cells because the permutation removes almost nothing. Measured on the training split, a teacher's relative tail has cosine similarity 0.994–0.996 to its own mean tail, with a minimum of 0.985. In a Fourier solution the logit at answer j depends on $(a + b - j)$, so in coordinates relative to the correct answer the tail is close to a fixed function of the offset and varies little across inputs. The shuffle is uninformative. It does not test whether the tail has to match the input. The identity is carried by the shape of that relative tail, which is a direct readout of the frequency spectrum.

Predicting a new signature by remapping the tail. Each swap above takes the tail from a second trained teacher. Here the tail is a rearrangement of one teacher's own tail, not a copy of another teacher's. Let c be the correct answer and r an invertible multiplier modulo p. At offset o from c, the target puts the source teacher's probability from offset $r\,o \bmod p$. Offset 0 and the two non-residue tokens stay as they were. The map only reassigns existing tail probabilities, so the correct-answer mass, the tail entropy, and a correct argmax are unchanged. On the real targets those values are 4.167 nats and mass 0.352 for teacher 4 at every r, and 4.084 nats and mass 0.292 for teacher 8. The map $o \mapsto ro \bmod p$ is an automorphism of the additive group $\mathbb{Z}_p$, which is why the spectrum stays sparse: an arbitrary entropy-preserving permutation would spread the power over every frequency. A Fourier component at frequency f moves to $rf \bmod p$, then folds to $\min(rf \bmod p,\ p - rf \bmod p)$ in $1, \ldots, (p-1)/2$. Unit tests check that property on ideal profiles and on full distributions (`tests/test_tail_remap.py`). The hypothesis is that the student's input-embedding key set becomes this predicted set $\{rf\}$, which no trained teacher in this paper computes through. The prediction uses the teacher's input-embedding key set under that same folding. Source and predicted sets: teacher 4, $\{10, 16, 19, 20, 28, 30, 32, 37, 38\}$; r=4: $\{15, 17, 21, 23, 31, 33, 40, 42, 46\}$; r=10: $\{3, 4, 6, 8, 9, 11, 18, 29, 34\}$. Teacher 8, $\{12, 24, 31, 33, 45, 46, 48\}$; r=4: $\{1, 2, 10, 14, 27, 35, 48\}$; r=10: $\{5, 19, 23, 25, 35, 39, 46\}$.

Before any student trained, the run manifest recorded the teachers' relative-logit spectra on the training split. The logit key sets are $\{16, 19, 20, 30, 32, 37, 38\}$ for teacher 4 and $\{12, 24, 31, 45, 46, 48\}$ for teacher 8. They match the embedding key sets at overlap 0.78 and 0.86. The match is close but not exact. Each logit set misses the embedding set's lowest-share frequencies. The two multipliers were chosen from the teachers alone, with no student data. Among $r \in 2..p-2$, r=4 is the smallest value whose predicted sets overlap the original sets least, summed over both teachers. r=10 is then the smallest value that adds the least overlap with those first predicted sets. Teacher 4's predicted sets are disjoint from its original set. Teacher 8's each share one frequency with it (48 under r=4, 46 under r=10), so those exclusive parts are ablated separately. A one-seed pilot, native and r=4, ran first and is included. The grid is $r \in \{1, 4, 10\}$ on both teachers, three seeds, using the swap experiment's training configuration and matched initialization per seed (Table 9).

Remapped students match their targets at held-out accuracy 1.000. KL to the target is 0.0017–0.0050, against 0.0007–0.0022 for native controls (Table 9). That endpoint is $\mathrm{KL}(\text{target} \,\|\, \text{student})$ over the full vocabulary at T=4, averaged over held-out inputs. Median relative-tail KL to the target is 0.002–0.006.

Every remapped student acquires the predicted set. Overlap with $\{rf\}$ is 0.86–1.00 and with the source set 0.00–0.08, the nonzero values being teacher 8's one shared frequency. Embedding power on the predicted set is 82–96%, against 1–12% on the original.

The predicted set is load-bearing. Removing it costs 0.94–0.99 of held-out accuracy in all twelve students. Removing only the part that is not in the original set costs 0.85–0.99. Removing the original set costs 0.00–0.01, and its exclusive part costs 0.00. Size-matched random sets cost 0.00 when they avoid both sets (4 draws) and at most 0.28 when they intersect one (56 draws). At the causal layer the probe decodes the predicted frequencies at R^2 0.77–0.99 and the original ones at 0.00–0.71. Native students decode their own at 0.79–0.99.

On the same checkpoints, a second test keeps only one frequency subspace of the residue input embeddings: the token mean plus the cos/sin planes of that set. Everything else in those embeddings

is projected out. All other weights stay as they are, including the non-residue rows and the output projection (Table 10). Keeping every frequency reproduces the clean logits to within 3.3×10^{-4}. Keeping only the predicted set leaves held-out accuracy at 1.000 in all twelve remapped students. KL to the target rises from 0.002–0.005 to 0.003–0.067. Keeping only the original set drops accuracy to 0.010–0.013 for teacher 4's students, chance being 0.010, and to 0.084–0.098 for teacher 8's, whose original set shares one frequency with the predicted set. For the native controls the two sets coincide, and the two restricted models agree.

Retaining the predicted input subspace preserves held-out accuracy in the trained network. Removing it sharply reduces accuracy. Retention does not preserve the full target distribution. A specified edit of the target's Fourier content selects a specified causal frequency set, with no model in this paper that already computes through that set. The intervention edits teacher-derived probabilities within one algorithm family. It says nothing about arbitrary data or about transferring a different algorithm. The manifest and result JSON are under `runs/stage13_tail_remap`. The design was committed before the runs (`stage13_design.md` in the artifact's preregistration folder).

teacher	r	KL to target	overlap		power		damage		probe R^2
			orig.	pred.	orig.	pred.	orig.	pred.	pred.
4	1	0.0007–0.0016	0.78–1.00	0.78–1.00	0.91–0.96	0.91–0.96	0.95–0.98	0.95–0.98	0.79–0.98
4	4	0.0027–0.0028	0.00	1.00	0.01	0.85–0.92	0.00	0.94–0.98	0.77–0.99
4	10	0.0017–0.0032	0.00	0.90–1.00	0.01–0.02	0.84–0.92	0.00	0.95–0.99	0.99
8	1	0.0008–0.0022	0.86	0.86	0.94–0.98	0.94–0.98	0.94–0.99	0.94–0.99	0.99
8	4	0.0020–0.0042	0.07–0.08	0.86–1.00	0.08–0.10	0.82–0.96	0.00	0.94–0.99	0.99
8	10	0.0031–0.0050	0.08	0.86–1.00	0.11–0.12	0.88–0.93	0.00–0.01	0.94–0.97	0.84–0.99

Table 9: Students trained on the source teacher's tail with offsets remapped by r (r=1 native), three seeds each, ranges over seeds; held-out accuracy is 1.000 for all eighteen. Overlap, embedding power share, input-embedding ablation damage and causal-layer probe R^2 are against the teacher's original key set and the predicted set $\{rf\}$; for r=1 the two coincide.

teacher	r	clean KL	keep predicted set only		keep original set only	
			acc.	KL	acc.	KL
4	1	0.0007–0.0016	1.000	0.006–0.032	1.000	0.006–0.032
4	4	0.0027–0.0028	1.000	0.018–0.038	0.011–0.013	1.360–1.556
4	10	0.0017–0.0032	1.000	0.003–0.035	0.010–0.013	1.280–1.384
8	1	0.0008–0.0022	1.000	0.003–0.021	1.000	0.003–0.021
8	4	0.0020–0.0042	1.000	0.006–0.067	0.088–0.098	0.977–1.053
8	10	0.0031–0.0050	1.000	0.015–0.022	0.084–0.095	1.057–1.075

Table 10: Keeping only one frequency subspace of the residue input embeddings in the remapped students, ranges over three seeds. KL is the endpoint to the student's own target; accuracy is held-out. For r=1 the predicted and original sets coincide.

Switching to the remapped target at different ages. The remapped students above start from a random initialization and acquire the predicted set. That does not say whether a student that already depends on its original set will do the same when given that remapped tail. We continue six native students from two ages onto the remapped target, and onto the original teacher as a control.

The young checkpoint is the first at which original-set ablation damage stays at or above 0.8. For all six trajectories (teachers 4 and 8, seeds 100–102) that checkpoint is step 500. The mature checkpoint is the $30{,}000$-step endpoint of the same native run. Each continuation is $30{,}000$ steps. The optimizer is reset and uses the same constant learning rate as the remapped students, so the mature branch is not running under a decayed schedule. The design was committed before the runs (`stage14_design.md`).

d_R, d_P and d_U are the held-out accuracy lost when the original set, the predicted set, or their union is removed from the residue input embeddings. Teacher 8's predicted set shares frequency 48 with its original set, so a student on the original set already has predicted-set overlap 0.08.

Young students were expected to re-select the predicted set. Mature students were expected to match the remapped outputs without switching the set they already use. Five of the six young students re-select it. Teacher 4 seed 102 fits the remapped output without either set becoming load-bearing on its own; removing their union still costs 0.98, so the computation stays inside the two sets jointly. None of the six mature students re-selects it. Five of them reach KL around 0.002, in the same range as the young branches, while remaining dependent on the original set or on a mixture that still includes it. The mixtures are partial adaptation, not rigidity: teacher 8 seed 100 loses 0.58 to predicted-set removal with 0.47 of its embedding power on that set, and teacher 4 seeds 100 and 101 drop their original-set damage to 0.42–0.46 while collapsing only under removal of the union. Teacher 4 seed 102 again fails to match (KL 0.032) and keeps the original set. In every pair, predicted-set overlap and predicted-set ablation damage are higher after the young switch than after the mature switch. Every native continuation stays on the original set (Table 11).

Two ages do not establish a critical period. The grid's result JSON is under `runs/stage14_switch_grid` and the one-seed pilot's under `runs/stage14_switch_time`.

teacher	seed	switch step	target	KL to target	overlap pred.	d_R	d_P	d_U	power R/P
4	100	500	native	0.0011	0.00	0.95	0.00	0.99	0.93/0.07
4	100	500	remapped	0.0017	1.00	0.00	0.97	0.98	0.04/0.87
4	100	30,000	native	0.0017	0.00	0.98	0.00	0.99	0.99/0.01
4	100	30,000	remapped	0.0020	0.08	0.42	0.00	0.98	0.70/0.21
4	101	500	native	0.0026	0.00	0.92	0.00	0.99	0.86/0.13
4	101	500	remapped	0.0028	1.00	0.00	0.96	0.96	0.01/0.87
4	101	30,000	native	0.0014	0.00	0.95	0.00	0.99	0.92/0.08
4	101	30,000	remapped	0.0018	0.08	0.46	0.01	0.97	0.71/0.22
4	102	500	native	0.0014	0.00	0.97	0.00	0.99	0.94/0.05
4	102	500	remapped	0.0016	0.38	0.00	0.16	0.98	0.38/0.59
4	102	30,000	native	0.0004	0.00	0.97	0.00	0.98	0.91/0.07
4	102	30,000	remapped	0.0321	0.00	0.99	0.00	0.99	0.93/0.01
8	100	500	native	0.0017	0.08	0.95	0.00	0.94	0.97/0.11
8	100	500	remapped	0.0034	0.86	0.00	0.94	0.95	0.13/0.94
8	100	30,000	native	0.0012	0.08	0.97	0.00	0.97	0.97/0.12
8	100	30,000	remapped	0.0029	0.20	0.84	0.58	0.98	0.63/0.47
8	101	500	native	0.0011	0.08	0.96	0.00	0.93	0.92/0.11
8	101	500	remapped	0.0018	0.86	0.00	0.98	0.99	0.19/0.88
8	101	30,000	native	0.0013	0.08	0.95	0.00	0.96	0.98/0.13
8	101	30,000	remapped	0.0027	0.09	0.96	0.12	0.99	0.87/0.25
8	102	500	native	0.0007	0.08	0.96	0.00	0.98	0.96/0.13
8	102	500	remapped	0.0018	0.75	0.00	0.97	0.97	0.18/0.87
8	102	30,000	native	0.0010	0.08	0.98	0.00	0.98	0.95/0.17
8	102	30,000	remapped	0.0019	0.09	0.98	0.00	0.98	0.92/0.20

Table 11: Young and mature continuations onto the native and remapped targets, $30{,}000$ steps each. The remapped target uses $r{=}4$. KL is to the continuation's own target. Power R/P is the embedding share on the original and predicted sets. Held-out accuracy is 1.000 at every listed point.

Causal dependence in the swap students. The original tail-swap experiment of Table 8 recorded overlaps but did not retain weights, so the transplanted set's causal role in the swap students themselves was not tested at the time. We retrained the own and swap cells of both orientations with the same configuration, seed for seed, kept the weights, and ablated three sets from each student's residue embeddings: the tail donor's key set, the mass donor's, and five size-matched random sets drawn from all frequencies (Table 12; the retrained key sets match the original records to within one frequency). Ablating the tail donor's set costs 0.92–0.99 of held-out accuracy in every swap student. Ablating the mass donor's set costs 0.00. Random sets of the same size cost a median of 0.00. The draws that cost more include key frequencies (at most 0.30). The own-tail students show the same dependence on their single teacher's set (0.91–0.99). The tail-donor set is therefore causally load-bearing in the from-scratch hybrids, not only present in their embedding spectrum. The small

models keep separate input-embedding and output-projection matrices, so this intervention, like every frequency ablation in the paper, acts on the input side only.

cell	mass	overlap		ablation damage		
	donor	tail donor	mass donor	tail-donor set	mass-donor set	random (median, max)
own tail	4	0.78–1.00	0.78–1.00	0.91–0.99	0.91–0.99	0.00, 0.37
own tail	8	0.86	0.86	0.94–0.99	0.94–0.99	0.00, 0.38
swapped tail	4	0.75–0.86	0.00	0.92–0.99	0.00	0.00, 0.30
swapped tail	8	0.89–1.00	0.00	0.97–0.99	0.00	0.00, 0.27

Table 12: Ablation in the retrained tail-swap students, five per cell, ranges over students; held-out accuracy is 1.000 for all twenty. Damage is held-out accuracy lost when that set is removed from the residue input embeddings.

Retargeting an established student. The swap students above start from a random initialization and acquire the tail donor's set. That does not say whether the same swapped tail can replace a frequency set that is already in place. We continue the $T{=}4$ students of teachers 4 and 8 from Section 4, three per teacher. From each student we run three continuations of 30,000 steps: gold labels, the student's own teacher, and the student's own correct-answer mass with the other teacher's tail. After the swapped-tail run, a further 10,000 steps of gold labels test whether the imitated tail lasts without tail supervision.

An established student was expected to emit the donor tail and keep the set it already uses. Under the swapped tail the students match the donor's outputs (KL 0.001–0.004). Overlap with the donor's set stays at 0.00–0.10, below the null 95th percentile of 0.25. Ablating the donor's set costs nothing in five of six students and 0.32 in one. Ablating the set they already use still costs 0.88–0.99 (Table 13). Once gold labels resume, the imitated tail drifts away from both teachers, and that set remains load-bearing. The own-teacher continuation keeps both the output and the set. Gold labels alone let the output drift while the set persists, with some erosion of overlap.

Fitting another teacher's tail does not replace the set already in use under this training budget. From a random start, the same target selects the donor set. The rest of the computation is untested; in particular the output projection, where the pretrained-scale signature concentrates (Appendix Q), was not analyzed in these students, and a static relative tail could be fitted there alone. The result JSON is under `runs/stage12_retarget_v2`. A first run with only the internal measurements is under `runs/stage12_retarget` and agrees throughout.

res.	arm, point	ov. res.	ov. donor	dmg res.	dmg donor	tail KL res. / donor
4	start	0.78–1.00	0.00	0.93–0.95	0.00	0.001–0.002 / 1.164–1.168
4	hard, 30k	0.67–1.00	0.00	0.94–0.98	0.00	0.180–0.298 / 0.614–1.244
4	own, 30k	0.78–0.89	0.00	0.92–0.97	0.00	0.000–0.002 / 1.162–1.164
4	swap, 30k	0.30–0.78	0.00–0.10	0.94–0.99	0.00–0.32	1.169–1.182 / 0.002–0.003
4	swap + 10k hard	0.40–0.67	0.00–0.09	0.94–0.99	0.00	0.365–0.540 / 0.452–0.474
8	start	0.86	0.00	0.93–0.96	0.00	0.001 / 1.180–1.187
8	hard, 30k	0.75–0.88	0.00	0.72–0.92	0.00	0.011–0.430 / 0.702–1.207
8	own, 30k	0.86	0.00	0.94–0.98	0.00	0.001–0.002 / 1.179–1.186
8	swap, 30k	0.71–0.86	0.00–0.07	0.88–0.98	0.00	1.135–1.154 / 0.001–0.004
8	swap + 10k hard	0.57–0.71	0.00	0.96–0.99	0.00	0.468–0.492 / 0.380–0.449

Table 13: Ranges are over three students. Overlap (ov.) and damage (dmg) are against the set the student already uses (res., the resident set) and against the donor's. Tail KL is the median held-out KL from the student's relative tail to each teacher at $T{=}4$. Held-out accuracy is 1.000 at every listed point.

I Optimization controls for capacity and KL direction

The capacity and KL-direction results could be statements about how far each condition trained. If a width-32 student is still descending when training stops, "capacity limits fidelity" and "this cell was undertrained" are the same observation. To understand this we read the recorded KD loss

back out of every run (`scripts/convergence_controls.py`). Remaining descent is the mean loss over the quarter of training before the final tenth, minus the mean over the final tenth, as a fraction of the former; a negative value means the final tenth sat slightly above the preceding quarter. Means replace single steps because per-step values on a full-batch objective oscillate enough to give spurious readings.

condition	final KD loss	remaining descent	own overlap
width 32	0.3167	1.9%	0.583 (0.42–0.67)
width 64	0.0237	0.2%	0.814 (0.75–0.86)
width 128	0.0179	−2.4%	0.857
width 256	0.0167	5.6%	0.886 (0.86–1.00)

Table 14: Capacity cells, all at $T{=}4$ so the losses are comparable. Own overlap is the mean over five seeds with the seed range in parentheses; a single value means every seed agreed.

Within the tested training schedule, width-32 students plateau at substantially higher KD loss and lower fingerprint overlap than wider students. Width 32 settles at a KD loss nineteen times that of width 256 with under 2% of its final-quarter descent remaining. The recorded loss curves show little late improvement, but do not exclude gains from another optimizer or longer training.

The KL-direction comparison needs more care, because forward and reverse KL minimize different quantities and their loss values are not comparable. Rescored on one common metric, the forward KL from teacher to student, the reverse-KL students sit *further* from the teacher (0.0039 against 0.0017) while carrying *higher* fingerprint overlap (0.456 against 0.205). Higher overlap under reverse KL is therefore not explained by a lower forward-KL error. The reverse-KL students have higher error on that common metric. This comparison does not exclude other effects of the optimization trajectory.

J SIGNATURES AND CAPABILITY BY CHANNEL

	channel / regime	information in signal	output signature	measured internal signature	capability
1	logit, sharp, $T{=}1$, fresh student	teacher tail present, strongly down-weighted	not separately measured	no reliable donor-signature match	not tested
2	logit, soft forward KL, fresh student	soft, teacher-specific	own teacher's tail (Table 13, start row)	tail donor's causal frequency set, graded	not tested
3	logit, reverse KL, $T{=}1$, fresh student	sharp $T{=}1$ target, reverse direction	not separately measured	partial teacher-specific overlap; causal ablation not run	not tested
4	logit, re-weighted $T{=}1$ tail, fresh student	teacher tail at raised weight	not separately measured	donor's causal frequency set	not tested
4a	logit, remapped tail, fresh student	transformed teacher-derived probabilities (offsets permuted by r)	matched (KL to target 0.002–0.005)	predicted set $\{rf\}$, causal	not tested
5	logit, swapped tail, established student	donor tail at $T{=}4$	donor's, fades under labels	resident-set dependence persists	unchanged accuracy
6	logit, automaton, forward KL	softened full distribution, 0.007 nats apart	not separately measured	below CKA match threshold	not tested
7	logit, automaton, re-weighted tail	conditional tails 0.2–0.5 nats apart	own teacher's, matched	below CKA match threshold	not tested
8	hidden-state	full residual stream	not separately measured	near-complete, partly by construction	not tested
9	trace (canonical state)	demonstrated steps	no teacher identity	procedure, causally verified	length generalization
10	trace (procedure demos)	demonstrated steps	—	composition not generalized	raises pass@k on two-hop queries, by narrowing
11	trace (knowledge)	new facts	—	—	expands pass@k on facts
12	GRPO	verifiable reward	—	not measured	Eval-P pass@k 0.00–0.03; little change

Table 15: Rows 1–8 leave capability untested: modular addition and the automaton are at ceiling, so there is no headroom to expand. They differ in what they install, which behavior cannot separate; rows 5 and 7 match the donor's output tail without meeting the corresponding donor-specific internal match criterion. Rows 9–12 assess capability. Row 9 is the automaton trace channel; rows 10–12 are the knowledge-graph factorial. Row 12 describes the setting tested here (Appendix D), not reinforcement learning in general.

K HOW MUCH TEACHER IDENTITY THE TRACES CONTAIN

Section 7 finds no own-teacher preference in trace students. That is informative only if the traces carry teacher identity, so we measure how much they carry.

Both parity-machine teachers (`decide_late` in the code) were rolled out greedily over identical symbol sequences, 2048 per length. Where two teachers emit the same state sequence, no procedure can tell them apart; where they differ, a lookup can. The best achievable accuracy of identifying

which teacher wrote a trace is therefore $\frac{1}{2}(1 + \Pr[\text{traces differ}])$ under an equal prior over the two teachers, a bound that needs no model and cannot be beaten.

length	traces differ	states differ	error A / B	best teacher ID
8	0.000	0.000	0.000 / 0.000	0.500
16	0.000	0.000	0.000 / 0.000	0.500
32	0.162	0.018	0.014 / 0.004	0.581
48	0.699	0.149	0.117 / 0.069	0.850

Table 16: Teacher identifiability in the traces. Lengths 8 and 16 are inside the training range. Lengths 32 and 48 are outside it.

Both teachers are perfect state trackers at training lengths, so their traces are identical to the gold traces and to each other. Teacher identification then sits at chance. Identity appears only out of distribution, where the teachers begin to err and diverge.

There are two consequences of this. First, the trace students' lack of own-teacher preference is arithmetic: they could not inherit what the traces do not contain. Second, it explains why the gold control is indistinguishable from the teacher-trace cells. At training lengths those traces are the same. The trace experiment establishes that correct canonical state traces teach the procedure they contain. Whether traces carrying teacher-specific content, such as errors, stylistic variation, or a different valid trace convention, would transfer identity is untested here.

L THE FREQUENCY-SET ABLATION

Fingerprint overlap says which frequencies appear in the embeddings. It does not say the model uses them to compute the answer. We remove a frequency set from the residue input embeddings by projecting out those cosine and sine directions, leave the rest of the network unchanged including the output projection, and measure held-out accuracy. Damage is the drop from the model's clean accuracy. Random controls are five sets, size-matched to the own-teacher key set, drawn from frequencies in neither teacher's set. The table reports the mean over those draws.

If the fingerprint is causal, removing a model's own set should collapse accuracy, and removing the other teacher's set or a random set should not. That holds for both teachers and for every $T{=}4$ student. $T{=}1$ students show no donor-specific pattern (Table 17). Removing one key frequency at a time does not reproduce the full-set damage. The worst single frequency costs teacher 8 0.267 and teacher 4 0.090, with mean damage 0.064 and 0.013 over key frequencies, against 0.92–0.93 for the full sets. The set is the causal unit. The result JSON is under `runs/stage6_ablation`.

model	trained via	teacher's set	other set	random sets
teacher 4	from scratch	0.92	0.00	0.00
teacher 8	from scratch	0.93	0.00	0.00
students of 4	fwd KL, $T{=}1$	0.00–0.13	0.00–0.21	0.02–0.23
students of 8	fwd KL, $T{=}1$	0.00–0.33	0.00	0.00–0.07
students of 4	fwd KL, $T{=}4$	0.93–0.96	0.00	0.00
students of 8	fwd KL, $T{=}4$	0.93–0.98	0.00	0.00

Table 17: Held-out accuracy lost when that set is removed from the residue input embeddings; ranges over five student seeds. For the teacher rows the teacher's set is the model's own key set; for student rows it is the distilling teacher's, which a $T{=}1$ student need not share.

M THE TEACHER-PAIR REPLICATION

The $T{=}1$ versus $T{=}4$ contrast in Section 4 uses one pair, teachers 4 and 8, chosen because their frequency sets are disjoint. Five student seeds of that pair measure optimization variability, not whether other teachers behave the same way. We distill from ten further teachers. Five pairs were

fixed before any student ran (`stage7_preregistration.md`). For each of five overlap targets spanning the population range, the rule takes the nearest pair that does not reuse a teacher. Teachers 4, 8, and 12 are excluded. From each of the ten teachers we train three students at $T=1$ and three at $T=4$ through forward KL, sixty students in all. The predictions:

P1. At $T=4$, own-teacher overlap exceeds the null 95th percentile of 0.25 in a majority of the ten teachers, and exceeds other-teacher overlap.

P2. At $T=1$, own-teacher overlap sits inside the null for a majority of the ten teachers, with no own-versus-other preference.

P3. At $T=4$, the own-minus-other difference decreases as the two teachers' fingerprints overlap more.

P1 holds in every teacher and every seed. Own-teacher overlap at $T=4$ still ranges from 0.33 to 1.00. Softening transfers teacher-specific structure in every case. P2 fails on its second clause: six of the ten teachers sit inside the null at $T=1$, as predicted, but four transfer, and own-teacher overlap exceeds other-teacher overlap in eight of ten. P3 does not hold: own-minus-other does not vary with pair overlap at either temperature (Table 18).

Teacher output entropy, measured on the training split, was not preregistered. Section 4 attributes $T=1$ transfer to a teacher being naturally soft, on the evidence of teacher 12. Ten teachers make that checkable. The four that transfer at $T=1$ are those with entropy 0.0101 nats or above. The six at 0.0037 or below do not. The interval locates the change. It is not an estimated threshold. The result JSON is under `runs/stage7_pairs`.

				$T=1$		$T=4$	
teacher	entropy	partner	pair overlap	own	own−other	own	own−other
6	0.0000	0	0.00	0.00–0.08	-0.20	0.33–0.57	+0.36
0	0.0002	6	0.00	0.06–0.31	+0.03	0.73	+0.73
15	0.0012	11	0.23	0.00	-0.12	0.75	+0.46
7	0.0018	16	0.31	0.07–0.33	+0.09	0.88–1.00	+0.64
16	0.0030	7	0.31	0.12–0.42	+0.13	1.00	+0.69
1	0.0037	2	0.10	0.09–0.38	+0.03	0.91	+0.80
3	0.0101	18	0.18	1.00	+0.82	1.00	+0.82
18	0.0207	3	0.18	0.29–0.70	+0.35	1.00	+0.82
11	0.0478	15	0.23	0.78–0.89	+0.60	0.89	+0.64
2	0.0588	1	0.10	0.82–0.91	+0.74	0.73–0.91	+0.71

Table 18: The temperature contrast across ten teachers, ordered by the teacher's own output entropy in nats. Partner is the other teacher of the prespecified pair and pair overlap the two teachers' key-set Jaccard, the independent variable of P3. Ranges are over three student seeds; own−other is the mean difference. Held-out accuracy is 1.000 in all sixty students.

N THE LOGIT-CHANNEL CONTRAST AT 16.9M PARAMETERS

Every teacher in Section 4 has width 128, and the largest student is width 256. A reader can take the temperature contrast as a property of those models. We repeat it with 4-layer, width-512 teachers and students of 16.9M parameters, about thirty times larger. Design and predictions were committed before any student ran (`stage8_preregistration.md`).

Two teachers, seeds 0 and 3, reach held-out accuracy 1.000 with disjoint key-frequency sets. They pass the identity criteria: probe R^2 0.887 and 0.842 on their own key sets, interchange 1.000 with a 0.000 control, and own-set ablation damage 0.990 and 0.866 against 0.000 from the other teacher's set. The probe values sit below the width-128 pair's 0.972 and 0.986. Output entropies, measured before any student trained, are 0.0807 and 0.0064 nats. The independent-teacher null is re-estimated at this size from fifteen teachers (105 pairs): mean 0.082, 95th percentile 0.268, maximum 0.556. The preregistration fixed that expansion in advance: ten further teachers, giving fifteen and 105 pairs, with the 95th percentile replacing the five-teacher maximum of 0.231 once they existed. The

extra teachers were trained after the students under that rule, the null uses teacher fingerprints only, and every verdict below is the same against 0.231.

Three students per teacher at each temperature, forward KL, $60{,}000$ steps, matching the teacher architecture.

P1. At $T{=}4$, for both teachers, own-teacher overlap exceeds the independent-teacher null at this size and exceeds other-teacher overlap, in a majority of the six students.

P2. At $T{=}1$, teacher 0 (0.0807 nats) transfers, because its entropy is above that of every teacher that transferred at $T{=}1$ in Appendix M.

P3. At $T{=}1$, teacher 3 (0.0064 nats) has no prediction. Its entropy sits in the interval Appendix M could not resolve.

P1 holds in every student. P2 holds in all three students. Teacher 3 at $T{=}1$ transfers in two of three, so the transition at this size can extend below 0.0064 nats (Table 19). That location is a property of the setting, not a fixed cutoff in entropy.

Teacher size, student size, width, depth, training duration, and the pair all change together. The comparison shows the contrast at 16.9M parameters. It does not isolate scale. The result JSON is under `runs/stage8_students`. The teacher identity results are under `runs/stage8_teacher_battery`.

teacher	entropy	T	held-out acc.	own	other	own−other
0	0.0807	4	1.000	0.60–0.67	0.07	+0.55
3	0.0064	4	1.000	0.40–0.75	0.00–0.08	+0.58
0	0.0807	1	1.000	0.36–0.56	0.00–0.09	+0.42
3	0.0064	1	0.997–1.000	0.17–0.71	0.00	+0.43
null at this size (105 pairs): mean 0.082, p95 0.268, max 0.556						

Table 19: Own and other are a student's key-frequency overlaps with the two teachers; own−other is the mean difference over three students.

O A WITHIN-RUN CHECKPOINT STUDY OF TEACHER ENTROPY

Section 4 attributes $T{=}1$ transfer to the teacher's own softness. A planned manipulation would have set that softness by training duration and weight decay while holding the mechanism fixed. Calibration showed that entropy oscillates across training over three orders of magnitude with no trend, and that key-frequency sets drift within a run. Checkpoints therefore select moments rather than set entropy. The study was rewritten before the main runs as an observational $2{\times}2$ (`stage9_preregistration.md`). No causal reading attaches to it.

From each of four teacher runs (seeds 20–23) we take four checkpoints: low entropy (≤ 0.0037 nats) or high (≥ 0.0101), crossed with early (first third of training) or late (last third). A checkpoint enters only if it reaches held-out accuracy 0.999, passes A1, A2 and A3 in full, and overlaps the run's consolidated fingerprint by at least 0.6 Jaccard. All sixteen did. Three students per checkpoint at $T{=}1$, plus a $T{=}4$ positive control from each run's lowest- and highest-entropy checkpoints, give 72 students. Every comparison is against the same fixed other-teacher (seed 24).

We predicted three outcomes. High-entropy checkpoints should transfer more than low-entropy ones at $T{=}1$. Duration should not matter. The $T{=}4$ control should inherit from both the lowest- and highest-entropy checkpoints.

cell	teacher entropy	own	own−other	held-out acc.
low, early	0.0023–0.0034	0.06–0.73	+0.11	0.999–1.000
low, late	0.0030–0.0037	0.00–0.58	+0.12	1.000
high, early	0.0103–0.0113	0.55–1.00	+0.73	1.000
high, late	0.0102–0.0108	0.17–1.00	+0.43	1.000
T=4 control	0.0023–0.0113	0.73–1.00	+0.71	1.000

Table 20: Own is a student's key-frequency overlap with its checkpoint teacher; own−other is the mean difference against the fixed other-teacher.

The entropy comparison is positive in all four runs: high-entropy checkpoints exceed low-entropy ones in own-minus-other by $+0.34$ to $+0.68$, mean $+0.46$. Duration was predicted not to matter, and it does. The early checkpoints exceed the late ones by $+0.01$ to $+0.28$, mean $+0.15$, and the entropy difference is larger early (interaction $+0.31$, positive in all four runs). Early checkpoints overlap their run's consolidated fingerprint at 0.60–0.89 against 1.00 late by construction, so duration and degree of consolidation cannot be separated here. That confound was stated before analysis. The T=4 control inherits in every student.

Two limits apply. Selection is not manipulation, and a common cause could drive both the entropy fluctuation and transferability. Section 6 also shows that scalar entropy is not the operative variable, since label smoothing raises entropy without transferring anything. The study is consistent with the entropy account. Causation is untested here.

P Re-weighting an unchanged tail

Temperature changes the training target and the loss weight on its tail at the same time. At T=1, students of sharp teachers do not inherit the frequency signature. That failure has two possible causes: the tail has nothing to transfer, or the loss gives the tail almost no weight. This test holds the T=1 tail fixed and only changes λ. Forward KL at T=1 splits into a Bernoulli term on the correct-answer probability plus the conditional-tail KL weighted by the teacher's wrong-answer mass, between 10^{-4} and 10^{-3} for these teachers. A sharp teacher therefore barely trains its own tail. The loss is $L = \mathrm{KL}(\mathrm{Bern}(m_T) \,\|\, \mathrm{Bern}(m_S)) + \lambda \,\mathrm{KL}(q_T \,\|\, q_S)$. The tail q is computed by log-softmax over the non-answer logits, never by dividing probabilities by $1 - m$. We train three students from each of teachers 4 and 8 at $\lambda \in \{0.1, 1\}$, and a uniform tail at the same weights as a content control.

If the tail already carries the signature, re-weighted students should inherit the donor set. A uniform tail at the same λ should not. Every re-weighted student inherits its donor's key set and depends on it causally (Table 21). λ=0.1 already suffices. Five of six uniform-tail students at λ=0.1 reach held-out accuracy 0.998–1.000 with overlap 0.00 against both teachers. The sixth (teacher 4, seed 102) stops at accuracy 0.49 and still shows no donor set. The λ=1 uniform arm also transfers nothing, but destabilizes training in two seeds, which is why the main text reports the λ=0.1 comparison. Plain T=1 students sit inside the null on the same measurements (Table 17). The sharp teacher's tail already carries the signature. Standard forward KL under-weights it.

teacher	arm	acc.	overlap own	overlap other	damage own	damage other
4	λ=0.1	1.000	0.78–0.89	0.00	0.83–0.92	0.00
4	λ=1	1.000	0.78–0.89	0.00	0.76–0.88	0.00
4	uniform, λ=0.1	0.490–1.000	0.00	0.00	0.00–0.02	0.00–0.02
4	uniform, λ=1	0.495–1.000	0.00–0.09	0.00–0.16	0.00–0.11	0.00–0.10
8	λ=0.1	1.000	0.86	0.00	0.47–0.90	0.00
8	λ=1	1.000	0.71–0.86	0.00	0.78–0.90	0.00
8	uniform, λ=0.1	0.998–0.999	0.00	0.00	0.00	0.00
8	uniform, λ=1	0.995–1.000	0.00–0.15	0.00–0.06	-0.00	-0.00

Table 21: Re-weighting the unchanged T=1 tail. Overlap is the key-frequency Jaccard against each teacher (independent-teacher null p95: 0.25); damage is held-out accuracy lost when a teacher's key set is ablated from the student.

The same re-weighting on the automaton. The automaton teachers' softened full distributions differ by only 0.0066 nats at $T{=}4$. The temperature sweep does not say whether the output carries a teacher-specific tail. We apply the same loss to teachers 0 and 1. Their $T{=}1$ conditional tails differ by 0.21 and 0.50 nats. From each teacher we train three students at $\lambda \in \{0.1, 1\}$, and three on a uniform tail at $\lambda{=}1$, for $20{,}000$ steps.

Re-weighted students should match their teacher's tail. They should still fail the own-teacher CKA threshold if that geometry is not in the output. They match the teacher's tail and sit far from the other teacher's (Table 22). Accuracy is 0.997–1.000 and interchange is 0.999. The uniform control lands far from both. The output side is also far less constraining here than in modular addition: a four-state answer leaves a three-class tail with two degrees of freedom per input, against 95 for the 96-class residue tail, so matching it fixes little of a 128-dimensional hidden state and many geometries produce the same tail. No re-weighted student meets the CKA threshold (null 95th percentile 0.86). Some seeds sit closer to their own teacher in raw CKA, still below that threshold. Teacher 0's students sit closer to teacher 1. The result JSON is under `runs/stage11_decoupled` and `runs/stage11b_dfa_v3`.

teacher	arm	acc.	tail KL own	tail KL other	CKA own	CKA other	interch.
0	$\lambda{=}0.1$	0.997–0.998	0.0005–0.0007	0.207–0.214	0.26–0.38	0.43–0.62	0.999
0	$\lambda{=}1$	0.997–1.000	0.0002–0.0003	0.211–0.213	0.18–0.28	0.33–0.49	0.999
0	uniform, $\lambda{=}1$	0.997–1.000	1.2461–1.2492	0.109–0.110	0.18–0.34	0.33–0.87	0.999
1	$\lambda{=}0.1$	0.998	0.0005–0.0009	0.483–0.537	0.10–0.69	0.10–0.25	0.999
1	$\lambda{=}1$	0.997–0.998	0.0003–0.0012	0.458–0.500	0.21–0.76	0.07–0.29	0.999
1	uniform, $\lambda{=}1$	0.999–1.000	0.1087–0.1103	1.245–1.252	0.32–0.72	0.17–0.39	0.999

Table 22: Re-weighting the automaton teachers' $T{=}1$ tails. Tail KL is the median held-out conditional-tail KL from the student to each teacher; CKA is against each teacher's best layer, independent-teacher null 95th percentile 0.86.

Q THE PRETRAINED-SCALE EXPORT IN FULL

The constructed-task results use small transformers whose mechanisms we can check. This appendix asks whether the tail-versus-mass finding holds for pretrained models on GSM8K, where we cannot check the mechanism. Before any student trains, the two teachers must differ by more than retraining noise, or the output channel has nothing teacher-specific to carry.

We preregistered an export to Qwen2.5-Math-7B, with a 1.5B run of the same gate first (`stage5_preregistration.md`). The 7B arm never ran; every teacher below is the 1.5B model. The design: LoRA fine-tune two teachers from the shared base on disjoint halves of GSM8K train, then distill students on hard labels, soft KL, and traces. The stop condition is computed from the teachers alone. $D(X, Y)$ is the median per-token $\mathrm{KL}(X \| Y)$ on solution tokens of the first 512 GSM8K test problems, teacher-forced on their reference solutions. Teacher A trains on one half of GSM8K train (3,736 examples), teacher B on the other (3,737), and $A2$ on A's half under a different seed. The gate requires $D(A, B) \geq 2\, D(A, A2)$ and $D(A, B) \geq 0.10\, D(A, \mathrm{base})$ at some temperature in $\{1, 2, 4\}$.

The preregistered LoRA recipe fails at every tested temperature: the seed-noise ratio is 0.98–1.11 against a required 2 (`runs/stage5/pilot15/gate.json`). Two teachers trained on disjoint data differ no more than two runs on identical data. No students were trained from that recipe. The ratio-to-base condition alone would have passed at $T \geq 2$. The seed-noise null catches it. The failure is scoped to short LoRA fine-tuning from one pretrained base. The rest of this appendix keeps the gate, probe set, split, and thresholds as preregistered and varies only the teacher-training recipe.

Five recipes share the halves and the seed assignment. The pilot is LoRA rank 16, $\alpha{=}32$, on every attention and MLP projection, learning rate 10^{-4}, 300 steps of batch 4 at maximum length 640. The second raises the rank to 256. The third trains all weights for the same 300 steps. The last two train all weights for 1,200 and 4,800 steps and give teacher B a divergent curriculum: B's half is sorted by solution length, and each batch samples from a window that starts at the short end and widens to

the full half by the midpoint of training, while A and $A2$ sample uniformly. Teacher differentiation beyond retraining noise therefore required this deliberate curriculum asymmetry, and the asymmetry is also the likely source of B's higher output entropy below.

If stronger training produces a recipe effect beyond seed noise, the seed-noise ratio should rise and eventually clear 2. It does, and only at 4,800 steps (Table 23). We then trained that recipe four times: twice on the original halves, and twice as independent pairs with resampled halves and fresh seeds. The seed ratio clears in all four. The base-ratio condition clears once (Table 24). Students were trained only from the draw that cleared both. At $T{=}1$ both $D(A, B)$ and $D(A, A2)$ collapse as the teachers approach determinism, so $T{=}2$ is the operative temperature.

recipe	steps	$D(A,B)/D(A,A2)$ $T{=}1$	$T{=}2$	$T{=}4$	$D(A,B)/D(A,\text{base})$ $T{=}2$	gate
LoRA $r{=}16$	300	1.02	1.08	1.11	0.121	fails
LoRA $r{=}256$	300	1.12	1.09	1.05	0.245	fails
full fine-tune	300	1.34	1.21	1.16	0.184	fails
full, curriculum	1200	1.55	1.46	1.36	0.237	fails
full, curriculum	4800	2.80	2.71	1.52	0.103	passes

Table 23: The recipe factorial against the preregistered gate. The 4,800-step row is the second of two trainings of that recipe.

training	seed ratio	base ratio	acc. A/B	gate
pair 1, training 1	2.55	0.096	–	fails
pair 1, training 2	2.71	0.103	0.655/0.655	passes
pair 2, new halves	2.25	0.070	0.685/0.655	fails
pair 3, new halves	2.12	0.098	0.705/0.695	fails

Table 24: Every training of the 4,800-step recipe, at $T{=}2$. Accuracy is greedy on 200 held-out problems; the first training's teachers were not retained for evaluation.

Teachers and students. A raw comparison of each student to teachers A and B is not a fair identity score if the teachers differ in entropy. On the first 200 GSM8K test problems under greedy decoding, A and B both answer 65.5%, and $A2$ and the base answer 69.5%. The pair the students see is accuracy-matched. Median per-token entropy at $T{=}2$ is 1.15 nats for A and 2.65 for B. Any student with flat tails therefore reads as B-like under a raw comparison, which is why Section 10 reports differences across arms.

Students are Qwen2.5-0.5B fine-tuned from the pretrained base for 600 steps of batch 4 at maximum length 640, AdamW at learning rate 10^{-5}, teacher-forced on GSM8K train reference solutions with loss on solution tokens only. Targets are built per position at $T{=}2$. Writing m for the mass donor's probability on the gold token and q for the tail donor's distribution over non-reference tokens, the target places m on the gold token and $(1 - m)\, q$ on the rest. Arms are named by mass donor then tail donor: AA and BB are each teacher's own distribution, and AB and BA exchange the tail. The uniform arm replaces q with the uniform distribution over non-reference tokens. The hard arm trains on gold labels. Three student seeds per arm.

The shuffled-tail control. A student might match a donor tail because those probabilities belong to this problem, or because the tail's shape is almost the same across inputs. Shuffling donor tails across problems tests that. The original arm rotated tails across positions in the batch. When the reference tokens differ, the built target no longer sums to one, so that arm does not isolate input correspondence. Its rows are kept as a superseded control. The corrected arm shuffles each donor tail across problems that share the recipient's reference token. Positions with no same-token partner in the batch, 27% of them, fall back to a uniform tail. The matched fraction is 0.73 in every seed.

Identity read-out. Identity is read on 256 held-out problems, teacher-forced on their reference solutions, at $T{=}2$. Δ is the median per-token KL from the student's non-reference tail to teacher A

minus the same to teacher B, a difference of medians. Positive Δ means the student's tail is closer to B's. If the tail donor sets the output signature, exchanging only the tail should move Δ toward the new donor, and exchanging only the mass should not. That is the pattern in Section 10. Table 25 lists every seed on the screening window. Student greedy accuracy on 100 test problems is 0.14–0.26 across all arms.

target	tail KL to A			tail KL to B		
gold labels only	0.942	0.956	0.949	0.863	0.859	0.869
teacher A, full distribution (AA)	0.629	0.627	0.617	0.702	0.716	0.706
teacher B, full distribution (BB)	0.626	0.647	0.638	0.538	0.614	0.551
A correct mass, B tail	0.666	0.651	0.651	0.627	0.624	0.611
B correct mass, A tail	0.576	0.615	0.585	0.644	0.743	0.663
A correct mass, uniform tail	0.896	0.912	0.911	0.773	0.779	0.776
A correct mass, B tail, rotated (superseded)	0.901	0.929	0.911	0.850	0.879	0.848
A correct mass, A tail, matched shuffle	0.686	0.683	0.668	0.700	0.682	0.675
A correct mass, B tail, matched shuffle	0.691	0.690	0.684	0.645	0.633	0.628

Table 25: Median per-token tail KL from each student to each teacher, all three seeds per arm, 256 held-out problems at $T{=}2$.

A scale-free read. The KL read-out is sensitive to tail entropy, and B is flatter than A. A uniform or otherwise flat tail can therefore look B-like without matching B's shape. Per position, over the non-reference tokens, we take the cosine between mean-centered, unit-normalized logits. A uniformly flatter tail does not move that metric. Writing $\Delta_{\cos}$ for a student's median cosine to A minus its median cosine to B, so that positive $\Delta_{\cos}$ is A-ward, the tail contrast and the mass contrast point to the same donors as under Δ, each under its own sign convention (Table 26). The uniform arm, which Δ placed B-ward through flatness, has the same cosine to both teachers, about 0.11 against 0.27–0.50 for every other arm.

What the matched-token shuffle retains. The corrected shuffle is now a usable control. The question is how much of the tail effect depends on which problem the tail came from. On the unscreened window, it retains 0.63 of the aligned A tail's gold-relative shift in $\Delta_{\cos}$ and 0.75 of the aligned B tail's. Most of the measured effect survives this partial shuffle. The uniform fallback alone would remove about 27% of the shift, since the uniform arm moves $\Delta_{\cos}$ by nothing, so the matched positions retain close to all of the aligned tail's effect. The remainder is input correspondence, or that fallback. The two are not separated here.

target	Δ (KL, nats)			$\Delta_{\cos}$		
gold labels	+0.068	+0.088	+0.067	+0.023	+0.022	+0.023
AA	-0.078	-0.091	-0.093	+0.047	+0.048	+0.048
BB	+0.077	+0.024	+0.079	+0.003	+0.003	+0.002
AB	+0.033	+0.024	+0.037	+0.007	+0.008	+0.007
BA	-0.077	-0.134	-0.087	+0.051	+0.050	+0.050
A mass, uniform tail	+0.116	+0.127	+0.127	-0.001	-0.004	-0.002
A mass, B tail, rotated (superseded)	+0.047	+0.042	+0.055	+0.016	+0.016	+0.015
A mass, A tail, matched shuffle	-0.019	-0.006	-0.013	+0.038	+0.038	+0.039
A mass, B tail, matched shuffle	+0.040	+0.047	+0.046	+0.010	+0.011	+0.011

Table 26: Per-arm identity read-outs on GSM8K test problems 512–767, which the teacher screen never saw: Δ is the KL read-out, tail KL to A minus to B; $\Delta_{\cos}$ the shape read-out; three seeds per arm. The rotated shuffle row is the superseded control kept for the record; the matched-token rows are the corrected control.

Where the signature lives. The tail donor sets which teacher's output signature the student acquires. This test asks where that signature sits in the 0.5B network. AA and AB students share data, order, and mass donor, and differ only in the tail, so a swap between them isolates the tail factor. The student ties its embedding to its output projection. Copying that matrix into a separate head leaves

the logits unchanged and splits the weights into the embedding (E), the decoder stack (D), and the output projection (H).

If the inherited signature is a readout effect, swapping the output projection should move most of the gap and swapping the embedding should move none. Swapping the output projection between same-seed AA and AB students moves 0.74–0.79 of the pure tail gap toward its donor, the decoder 0.18–0.20, the embedding nothing, and the two together the whole gap (Table 27). Same-arm, cross-seed swaps move at most 0.02. Teacher-forced token accuracy stays in a narrow band throughout.

from AB	combo	$\Delta_{\cos}$ by seed			gap moved	token acc.
AA reference	YYY	+0.0468	+0.0475	+0.0477	0.00	0.854–0.857
E	XYY	+0.0469	+0.0476	+0.0476	0.00	0.854–0.857
D	YXY	+0.0386	+0.0401	+0.0396	0.18–0.20	0.855–0.856
H	YYX	+0.0166	+0.0152	+0.0162	0.74–0.79	0.854–0.856
E+D	XXY	+0.0385	+0.0402	+0.0395	0.18–0.20	0.855–0.856
E+H	XYX	+0.0165	+0.0153	+0.0163	0.74–0.79	0.854–0.857
D+H	YXX	+0.0060	+0.0068	+0.0064	1.00	0.855–0.857
AB reference	XXX	+0.0059	+0.0069	+0.0065	1.00	0.855–0.857
same-arm null, AA	E	+0.0469			0.00	0.854
same-arm null, AB	E	+0.0059			0.00	0.857
same-arm null, AA	D	+0.0468			0.00	0.856
same-arm null, AB	D	+0.0069			-0.02	0.858
same-arm null, AA	H	+0.0465			0.01	0.853
same-arm null, AB	H	+0.0060			0.00	0.858

Table 27: Evaluated on GSM8K test problems 512–639. Each combo is E, D, H in that order, with Y = AA and X = AB. Gap moved is the fraction of that seed's AA–AB $\Delta_{\cos}$ gap. Nulls swap the same component across seeds of one arm.

The decoder share could be spread through the stack or concentrated near the readout. Depth-block transplants put it at the top: layers 18–23 carry 0.18–0.25 of the gap, the other quarters 0.05 or less (Table 28).

transplanted block	fraction of the gap moved toward the donor		
	$AB \to BA$	$BA \to AB$	same-arm null
layers 0–5	0.01–0.03	-0.00–0.01	0.01–0.03
layers 6–11	0.01–0.02	-0.00–0.01	0.00–0.01
layers 12–17	0.00–0.05	0.00–0.02	0.01
layers 18–23	0.20–0.22	0.18–0.25	0.00–0.01

Table 28: Depth-block transplants between the paired hybrid students, reported as the fraction of the $\Delta_{\cos}$ gap between the two students that the recipient moves toward the transplanted block's donor, ranges over three seeds, evaluated on GSM8K test problems 0–127.

The same last-quarter concentration holds for the pure-tail pair AA/AB on unscreened problems (GSM8K test 512–639): layers 18–23 move 0.13–0.17 of that gap. The BB/BA pair, which shares a mass donor and exchanges the tail, puts 0.16–0.20 of its gap in those layers. A readout that reproduces a teacher's tail is not evidence of an inherited algorithm. The localization may also be a property of the schedule: 600 steps at batch 4 is under a third of an epoch, the students reach only 0.14–0.26 greedy accuracy, and the output projection is the tied embedding, which receives a dense gradient from every logit at every position. Whether a longer distillation moves the signature deeper is untested.

Dose along the identity direction. Component swaps show where the gap lives. Interpolating the full weight difference asks whether that direction is enough to place the signature, including in a student that never saw either tail. Writing ΔW for the AB minus AA weight difference, a gold-labels student plus $\beta\,\Delta W$ moves $\Delta_{\cos}$ from the gold baseline to the AA endpoint at $\beta = -1$

and past the AB endpoint at $\beta = +1$ ($+0.047$–$+0.049$ and -0.006, against $+0.047$–$+0.048$ and $+0.006$–$+0.007$ for the AA and AB students). Token accuracy holds at $|\beta|=0.5$ and dips at $|\beta|=1$, once the signature has already moved.

A designated test on a gate-failing pair. The gated experiment trained students only from the draw that cleared both gate conditions. The third pair cleared the seed-ratio condition and missed the base-ratio condition by 0.002 (Table 24). The question is whether the tail still transfers when that cutoff is missed by that margin. After the gated experiment we trained gold, AA, AB, and matched-shuffle students on this pair, three seeds each. The clean tail contrast AB against AA is the same size as on the pair that cleared (Table 29). The matched shuffle retains 0.84 of the B tail's gold-relative shift.

target	Δ (KL, nats)			$\Delta_{\cos}$		
gold labels	+0.051	+0.068	+0.056	+0.005	+0.006	+0.006
AA	-0.073	-0.082	-0.071	+0.030	+0.029	+0.030
AB	+0.026	+0.023	+0.025	-0.012	-0.013	-0.012
A mass, B tail, matched shuffle	+0.050	+0.046	+0.050	-0.009	-0.010	-0.009

Table 29: Students of the gate-failing third pair on GSM8K test problems 512–767, three seeds per arm; columns as in Table 26.

The mass–tail interaction. Exchanging the tail under B's mass moves Δ farther than exchanging it under A's. That could be noise, or the mass donor could modulate how strongly the tail is trained. The per-problem paired interaction, the B-mass tail effect minus the A-mass tail effect, has median $+0.024$ to $+0.035$ nats across seeds, with problem-level bootstrap 95% intervals that exclude zero. Under plain forward KL the reference-token mass sets the tail's training weight, the coupling decoupled knowledge distillation removes (Zhao et al., 2022). That coupling is a plausible account of the modulation. It is not a demonstrated one.

The result JSON is under ten `runs/stage5*` directories: the LoRA pilot (`stage5`), the recipe factorial and the two further pairs (`_recipes`, `_recipes_pair2`, `_recipes_pair3`), the students (`_students`, `_students_pair3`), the untied surgery (`_untied`), and the transplants (`_transplant`, `_transplant_aa_ab`, `_transplant_bb_ba`).

R ATTENTION DEPENDENCE AFTER TAIL MATCHING

Section 4 shows that the tail selects which frequency set a fresh student uses. Frequency overlap is not the whole computation. This appendix tests whether matching a teacher's tail also transfers that teacher's dependence on its learned attention pattern.

The intervention replaces every attention row, in both layers and all four heads, by the uniform distribution over the keys the causal mask leaves available. It is applied at inference only. A second intervention, the mean pattern, replaces each row by that head's mean attention pattern over the evaluation inputs at the same query position. It removes the pattern's dependence on the input but keeps its learned positional shape, so the two together separate content-dependent addressing from positional routing. We screen the twenty modular-addition teachers of Section 3 under the uniform intervention. Two whose accuracy falls, seeds 0 and 6, are each paired with one that does not. We call a teacher whose accuracy falls D, for attention-dependent, and one that remains accurate R, for robust. The labels describe that response, not the Clock and Pizza algorithm identities of Zhong et al. (2023). Zhong et al.'s gradient-symmetricity and distance-irrelevance diagnostics do not separate these teachers. All twenty are near-symmetric (0.96–0.99). In most of them the answer-position logits depend little on the operand difference.

On their own held-out halves, seventeen teachers keep accuracy 0.999–1.000 under the intervention, seeds 0 and 6 fall to 0.566 and 0.370, and seed 7 sits between at 0.918 (Table 30 reports the four used on the shared window below). Teacher 6 is paired with the paper's verified teacher 4. Teacher 0 is paired with teacher 13, whose key set overlaps teacher 0's most among the seventeen (0.24). The pairs are $(6, 4)$ and $(0, 13)$.

For each pair we train three freshly initialized students in each of seven arms: DD, DR, RD, RR, D-uniform, R-uniform, and hard labels. The first letter is the correct-answer mass donor and the second the conditional-tail donor. The uniform arms replace the tail with a uniform distribution over the incorrect residues. Forty-two students in all. The experiment was not preregistered. The 0.98 accuracy threshold under the intervention was fixed from the teacher screen before any student trained. The arms follow the tail-swap factorial of Appendix H.

The predictions: students given D's tail should match that tail and acquire D's frequency structure, and they should inherit D's accuracy collapse under forced-uniform attention. Students given R's tail should remain accurate, as R does.

Training and scoring. Teachers and students share the architecture of the main experiments (Appendix A). Distilled students use the tail-swap objective of Appendix H. Targets are built per input at $T{=}4$ from the mass donor's correct-answer probability and the tail donor's relative tail. The student minimizes the T^2-scaled forward KL to that target on the mass donor's training half (4,704 operand pairs, full batch) for $30{,}000$ steps of AdamW at learning rate 10^{-3}, $\beta = (0.9, 0.98)$, weight decay 1.0, and 10 warmup steps. A student seed sets the initialization. The same three seeds are used in every arm, so arms are paired by initialization, and full-batch training leaves no data order to match. Hard-label students train with the teacher recipe of Table 3 on the split their own seed defines.

Tail KL is the median, over evaluation inputs, of the KL from the model's relative tail to the unmodified reference teacher's, both at $T{=}4$, over the 96 incorrect residues. The gold answer is excluded whether or not the intervened model answers correctly. Each teacher defines its own split, so a student trained on R's half has seen about half of D's held-out half. Distilled students and the teachers are therefore scored on one fixed window that both teachers held out (2,340 and 2,361 pairs in the two pairs), which no distilled student saw, before and after the intervention on identical inputs. Hard-label students' accuracies are on their own held-out half.

Teacher contrast. Under forced-uniform attention, the two D teachers fall to accuracy 0.379 and 0.567, and their conditional tails move 0.42 and 0.51 nats from their unmodified distributions. The R teachers keep accuracy 1.000 and their tails do not move at the reported precision. Under the mean pattern both D teachers keep accuracy 1.000 and their tails move by at most 0.002 nats (Table 30). What they depend on is therefore the learned positional shape of their attention, not its variation with the input; no model here needs content-dependent addressing.

pair	teacher	acc.	uniform attn		mean pattern	
			acc.	tail KL	acc.	tail KL
(6, 4)	D, seed 6	1.000	0.379	0.422	1.000	0.000
(6, 4)	R, seed 4	1.000	1.000	0.000	1.000	0.000
(0, 13)	D, seed 0	1.000	0.567	0.507	1.000	0.002
(0, 13)	R, seed 13	1.000	1.000	0.000	1.000	0.000

Table 30: The four teachers of the attention intervention, scored with attention forced uniform and with each head's mean pattern imposed. Tail KL is from the intervened teacher's relative tail to its own unmodified tail, median over the shared window at $T{=}4$.

Student results. All twelve students receiving D's tail, across both mass donors and both pairs, match that tail at KL 0.001–0.002 before the intervention. Their key-set overlap is 0.31–0.91 with D and 0.05–0.22 with R. Under forced-uniform attention their accuracy remains 0.982–1.000, while their tail KL to the unmodified donor rises to 0.06–0.21 nats; under the mean pattern every distilled student keeps accuracy at least 0.998. Matching the tail before the intervention therefore coexists with a different accuracy response afterward. The tail match itself does not survive the intervention, so the students' learned attention is not inert: what survives is the argmax, not the distribution (Table 31).

The twelve students receiving R's tail also match their donor at KL 0.001–0.002 and show donor-associated frequency structure: overlap 0.50–1.00 with R and 0.08–0.24 with D. Their tails also change under the intervention, by 0.03–0.24 nats. The ranges overlap, so we do not claim that the tail donor controls how much the tail moves.

Controls. All six hard-label students keep accuracy 1.000 under forced-uniform attention. Uniform-tail students keep 0.866–1.000, and four of twelve fall below the fixed 0.98 criterion, as does one RR student (0.960). Remaining accurate under the intervention is the common outcome in this training setup, and it is the architecture's default: seventeen of twenty hard-label teachers and all six hard-label students show it, and output-only supervision leaves the routing unconstrained, so a fresh student has no reason to leave that basin. It is not established for every student. Continuous scores and the 0.98 threshold are reported together in Table 31; 37 of the 42 students meet the criterion.

pair	arm	acc.	uniform-attn	tail KL to tail donor		overlap		pass
			acc.	normal	uniform attn	D	R	≥ 0.98
(6, 4)	DD	1.000	1.000	0.001	0.061–0.162	0.31–0.50	0.05–0.13	3/3
(6, 4)	DR	1.000	0.995–1.000	0.001–0.002	0.029–0.242	0.12–0.15	0.50–0.82	3/3
(6, 4)	RD	1.000	0.993–1.000	0.001–0.002	0.090–0.143	0.44–0.50	0.12–0.13	3/3
(6, 4)	RR	1.000	0.960–1.000	0.001–0.002	0.053–0.209	0.08–0.10	0.78–1.00	2/3
(6, 4)	D-uniform	0.998–1.000	0.885–1.000	–	–	0.00	0.00–0.13	2/3
(6, 4)	R-uniform	0.999–1.000	0.961–1.000	–	–	0.00–0.09	0.00–0.06	2/3
(6, 4)	hard labels	1.000	1.000	–	–	0.00–0.08	0.06–0.20	3/3
(0, 13)	DD	1.000	0.982–1.000	0.001–0.002	0.102–0.210	0.73–0.91	0.18–0.22	3/3
(0, 13)	DR	1.000	1.000	0.001–0.002	0.102–0.160	0.24	1.00	3/3
(0, 13)	RD	1.000	0.993–1.000	0.001–0.002	0.125–0.199	0.73	0.18	3/3
(0, 13)	RR	1.000	0.999–1.000	0.001–0.002	0.058–0.162	0.24	1.00	3/3
(0, 13)	D-uniform	0.999–1.000	0.866–0.998	–	–	0.00–0.12	0.00–0.12	2/3
(0, 13)	R-uniform	1.000	0.962–1.000	–	–	0.00–0.13	0.00–0.18	2/3
(0, 13)	hard labels	1.000	1.000	–	–	0.12–0.31	0.05–0.24	3/3

Table 31: Fresh students, three seeds per arm. DD, DR, RD, and RR name mass donor then tail donor; D is attention-dependent and R is robust. Ranges are over seeds. Tail KL is to the unmodified tail donor. Pass is held-out accuracy at least 0.98 under forced-uniform attention.

Scope. These results concern one intervention in one architecture and task family, replicated across two teacher pairs. Close output matching does not guarantee matching accuracy dependence on the learned attention pattern. Uniform attention still mixes value vectors. The intervention does not identify the student's full algorithm. It also does not show that learned attention is unused in the unmodified network. Frequency overlap alone does not establish causal use. In a follow-up evaluation of the saved checkpoints, we combined donor-frequency retention or removal with forced-uniform attention (Table 32). In pair $(0, 13)$, all six D-tail students depend on the donor's frequency set for high accuracy under both attention conditions. Retention gives accuracy 1.000 with normal attention and 0.937–1.000 with uniform attention, whereas removal reduces accuracy to 0.107–0.211 and 0.047–0.079, respectively. The donor teacher satisfies the same necessity criterion (removing its own set costs it 0.854 under normal attention) and remains impaired by uniform attention, scoring 0.567 with unmodified embeddings and 0.689 when only its own frequency set is retained. Thus, donor-frequency dependence can transfer without the teacher's accuracy dependence on its learned attention pattern.

The same conclusion is not established for pair $(6, 4)$. Removing teacher 6's threshold-defined key set leaves the teacher at accuracy 0.990. So this set is not necessary for its high accuracy. Students likewise retain high normal-attention accuracy after removal, while retaining only that set is insufficient for consistently high accuracy. We report both pairs. The stronger causal conclusion is supported for pair $(0, 13)$. These interventions do not identify the students' full algorithms.

pair	students	donor-frequency intervention	normal-attn acc.	uniform-attn acc.
(6, 4)	6	keep only donor set	0.725–1.000	0.223–0.822
(6, 4)	6	remove donor set	0.998–1.000	0.790–1.000
(0, 13)	6	keep only donor set	1.000	0.937–1.000
(0, 13)	6	remove donor set	0.107–0.211	0.047–0.079

Table 32: Joint frequency and attention interventions in D-tail students, three DD and three RD students per pair. Values are ranges over the same six students in each pair. Donor sets use the fixed embedding-power threshold. Retention keeps the donor-frequency subspace plus the embedding mean; removal projects out that subspace. Other weights, including the output projection, remain unchanged. Attention is either unmodified or forced uniform as defined above.

For pair $(0, 13)$, five size-matched random-removal controls leave student accuracy at 1.000 under normal attention and at least 0.917 under uniform attention. Retaining only random frequency sets gives accuracy at most 0.077 and 0.049, respectively. Retention therefore supports high answer accuracy specifically in the donor subspace. It does not imply preservation of the full output distribution.

Under `runs/stage15_attention` are the teacher screen (`teacher_screen.json`), the training records (`students/results.json`), the fixed-window scores (`intervention_6_4.json` and `intervention_0_13.json`), and the follow-up joint evaluation (`joint_intervention.json`). The code is the three `scripts/stage15_attention_*.py` files, one each for the screen, the students, and the scoring, and `scripts/stage15_joint_intervention.py` for the follow-up.